Causal Inference Methods in Air Pollution Research

Causal Inference Methods in Air Pollution Research

A dissertation presented

by

Georgia Papadogeorgou

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2018

Dissertation Advisors: Francesca Dominici & Corwin M. Zigler Georgia Papadogeorgou

Causal Inference Methods in Air Pollution Research

Abstract

While the air pollution concentrations in the United States continue to decrease, one important and politically charged question remains: Is long-term exposure to low levels of air pollution still harmful? Several approaches have been developed to estimate the relationship between exposure to particulate matter with diameter at most 2.5 micrometers ($PM_{2.5}$) and various health outcomes. However, none of these approaches account for the fact that different variables might act as confounders of the exposure response relationship at different exposure levels.

In chapter 1, we developed a Bayesian methodology for the estimation of the causal exposure-response curve for exposure to $PM_{2.5}$ on cardiovascular hospitalizations. This method allows for flexible estimation of the shape of the exposure-response relationship, and for differential confounding adjustment at different levels of the exposure. Moreover, it provides a principled way to identify the confounding importance of different predictors at different exposure levels.

Over the last few decades, there have been various regulations in the Unites States aiming to reduce emissions from power plants with the ultimate goal of reducing ambient air pollution concentrations and pollution-related hospitalizations. However, the effectiveness of these regulations has not been adequately studied. Since nitric oxide and nitrogen dioxides (NO_x) are important precursors of ozone formation, we focused on the comparative effectiveness of a class of NO_x emission reduction technologies against alternatives on ambient ozone concentrations.

In chapter 2, we developed causal inference methodology rooted in propensity score matching to adjust for unobserved spatial confounding, such as unmeasured weather and

atmospheric conditions. We showed that unobserved confounding by spatial variables is likely to be present, and that incorporating spatial proximity in the matching of treated units to control units returns effect estimates that are more in line with subject-matter knowledge.

In chapter 3, we addressed the issue of interference in the studies of air pollution regulations. The movement of emissions and air pollution leads to interference, since interventions that take place at one power plant can affect air pollution levels in the area surrounding other power plants. In more detail, assuming that the power plants can be clustered in groups within which there is interference but not across them, we defined new estimands for causal inference with interfering units that correspond to quantities of interest under realistic treatment allocation programs. Consistent estimators and asymptotic results were derived and were employed to quantify the comparative effectiveness of NO_x emission control technologies.

Contents

Acknowledgments

First of all, I would like to thank my dissertation advisors Francesca Dominici and Cory Zigler for their unlimited professional and personal support during my time at Harvard. You are exceptional researchers, mentors and people. I consider myself lucky to have studied, learnt, and grown with you. You have taught me so much about academia, you have allowed me to explore, supported all my ideas, and listened to my insights. I recognize that what's yet to come I owe greatly to you.

I would also like to thank Fabrizia Mealli from the University of Florence for teaching me so much about causal inference, and for all our conversations about research and personal life. You helped me grow tremendously as a researcher.

Furthermore, I would like to thank all of my professors at the Department of Mathematics of the University of Athens, but most importantly Prof. Dimitrios Varsos for growing my interest in mathematics, and for being by far the most inspiring teacher I have ever had. But also, I would like to thank Prof. Dimitris Cheliotis who urged me to apply to Ph.D. programs in the U.S., and without whom I would have never applied to Harvard. Thank you for all you've done for me, and for continuing to be an encouraging force in the department.

I would also like to thank my cousin Eleni for being my best friend and my partner in crime since the day I was born (because you're older than me!) and all my friends from back home for having to deal with me when I carried books everywhere and talked about mathematics too much. But also, all the special friends I've made in Boston, wherever you might all be now. Thank you for making this place feel like home, for being there for all the happy, sad and crazy moments. You made this journey so enjoyable!

I would especially like to thank Dr. Joseph Antonelli, who was only a Mr. when I met him. You have seen me at my best and you have seen me at my worst, you have been there when I missed home and when I felt that the world couldn't fit me. I am greatly grateful for your company, the last five years have been infinitely better because I had you by my side.

Lastly, I would like to thank my family, my parents Evi and Mpampis and my brother

Michalis. Μαμά, μπαμπά, Μιχάλη, ευχαριστώ. You made me strong and able to fight for what I want, you allowed me to be myself, you taught me how to love who I am and respect others around me. I don't know who I would be without you.

I would like to dedicate this dissertation to my grandparents Γεώργιο and Γεωργία Στραβοδήμου. You were the greatest role models a kid can have. I know how much you wish you were here for this. Γιαγιά, σ΄ ευχαριστώ για όλα.

To the loved ones that are here,
and to the ones that aren't.

Local confounding adjustment for estimating health effects at low air pollution levels

Georgia Papadogeorgou

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Francesca Dominici

Department of Biostatistics

Harvard Chan School of Public Health

1.1 Introduction

As air pollution levels decrease and air quality interventions become more costly, epidemiological evidence of the potential public health benefits of further reductions have become the subject of intense scrutiny. The literature on the harmful effects of air pollution is very extensive (Dominici et al., 2002; Eftim et al., 2008; Zeger et al., 2008; Zanobetti and Schwartz, 2007), but significant substantive and methodological gaps remain. In fact, there is a need for methods for causal inference that allow estimation of a flexible exposure response (ER) function coupled with robust methods for confounding adjustment.

Parametric and semi-parametric regression modelling approaches for ER estimation have been proposed in the literature in the context of clinical trials data (Babb et al., 1998), toxicology (Scholze et al., 2001), and air pollution research (Bell et al., 2006; Daniels et al., 2000; Dominici et al., 2002; Schwartz et al., 2002; Shi et al., 2016). Regression and semi-parametric modeling approaches for ER estimation such as Generalized Linear Models or Generalized Additive Models (Hastie and Tibshirani, 1986; Daniels et al., 2004; Shaddick et al., 2008; Shi et al., 2016; Dominici et al., 2002), generally make the following assumptions: 1) the same potential confounders are considered when estimating the health effects across all exposure levels (i.e. global confounding adjustment); 2) the set of potential confounders that are included into the regression model among a potentially large set of available covariates is specified a priori; 3) these pre-selected potential confounders are included into the model as linear or spline terms for confounding adjustment (i.e. parametric/semi-parametric adjustment for confounding bias); and 4) the shape of the ER function is modelled as a spline, a polynomial, or linear with a threshold.

In the causal inference literature, Hirano and Imbens (2004) introduced the generalized propensity score (GPS) in order to adjust for confounding when estimating the causal effects of a continuous exposure. More recently Kennedy et al. (2017) introduced a doubly robust approach for estimating the causal ER function. Although these approaches are really promising, they still rely on global confounding adjustment of pre-selected potential confounders, and do not provide guidance of the covariates' confounding importance at different exposure levels.

However, there is evidence that the relationship between exposure to air pollution (fine particulate matter $PM_{2.5}$) and health outcome (rate of hospitalization for cardiovascular diseases) might be confounded by a different set of covariates at the low exposure levels versus at the high exposure levels, as we will demonstrate in Section 1.2. We refer to this differential confounding across exposure levels as *local confounding*, since different sets of covariates confound the air pollution effects "locally". In simulation studies, we will demonstrate that when local confounding is present, common approaches for ER estimation lead to biased estimates. We argue that –especially in the context of estimating causal effects at low levels of exposure– local confounding adjustment is deemed necessary.

To target local confounding, one could model data separately at different exposure levels, including all potential confounders. However, even if the exposure levels with differential confounding were known, inclusion of all covariates in an outcome model could lead to inefficient estimation of causal effects at exposure levels with a small sample size. This is particularly true in our data application where the estimation of effects at low exposures coincides with a small number of observations at that level. Data driven methods to select the minimum necessary set of covariates to be included into an outcome model for estimation of causal effects have been proposed (Luna et al., 2011; Wang et al., 2012; Wilson and Reich, 2014), but to our knowledge, they have not been extended to the context of ER estimation and local confounding adjustment.

In addition, although parametric or semi-parametric modelling of the ER are attractive for their flexibility in identifying different shapes (see for example Dominici et al. (2002); Daniels et al. (2000); Scholze et al. (2001); Schwartz et al. (2002); Govindarajulu et al. (2009); Shaddick et al. (2008)), they have the drawback of heavy reliance on the model specification when extrapolating evidence on health effects at the very low exposure levels. For example, smooth functions do not allow a hockey stick shape of ER curve by construction, which is one of our key epidemiological questions.

In this paper, we introduce a Bayesian framework for the estimation of a causal ER curve termed LERCA (Local Exposure Response Confounding Adjustment). LERCA aims to overcome some of the challenges described above. We introduce the concept of *experiment configuration* which consists of $\boldsymbol{s} = (s_0, s_1, \ldots, s_{K+1})$, where $[s_{k-1}, s_k)$ denotes a spe-

cific range of exposure values. We use the term *experiment* for the hypothetical assignment of a unit to exposure value within $[s_{k-1}, s_k)$. Within each experiment, i.e. locally in the exposure range $[s_{k-1}, s_k)$, we assume that: 1) ER is linear; and 2) the potential confounders of the exposure-outcome relationship are unknown but measured. Importantly, the *experiment configuration* $\boldsymbol{s}$ is unknown and it will be estimated from the data. LERCA allows for local confounding adjustment, and provides guidance related to the observed covariates' importance as confounders or outcome predictors at different exposure levels, which is a key question in epidemiological studies of air pollution.

In Section 1.2 we introduce the dataset, and we illustrate the issue of differential confounding at the different exposure levels. In Section 1.3, we introduce the notation used and the assumptions on which LERCA relies. LERCA is presented in Section 1.4, and in Section 1.5 it is illustrated and compared to alternative methods in the presence of local or global confounding through extensive simulations. Finally, we apply LERCA to estimate the causal ER function of long term exposure to $PM_{2.5}$ on cardiovascular hospitalization rates in Section 1.6. Limitations and potential extensions are discussed in Section 1.7.

1.2 Data description and illustration of local confounding

We assemble a data set where the unit of the observation is the zip code i, with sample size $N = 5,362$. For each zip code, we calculate: 1) y_i defined as log hospitalization rate for cardiovascular diseases (codes ICD-9 390 to 459) among Medicare participants residing in the zip code i in the year 2013; 2) x_i defined as average exposure to air pollution obtained by averaging ambient $PM_{2.5}$ levels for the years 2011 and 2012 from EPA monitoring sites within 6 mile radius of zip code i's centroid. The values of x_i vary across zip codes from min $= 2.7$ to max $= 18.3 \mu g/m^3$. Figure 1.1 shows the locations of the zip codes' centroids and the corresponding exposure values for $PM_{2.5}$. Our final data set includes 27 potential confounders, with C_j denoting variable $j = 1, 2, \ldots, p$ and $p = 27$. These variables include demographics from the US Census Bureau, climate data from the Automated Surface Observing System of the National Oceanic and Atmospheric Administration, and covariate information from the Behavioral Risk Factor Surveillance System and the Dartmouth atlas

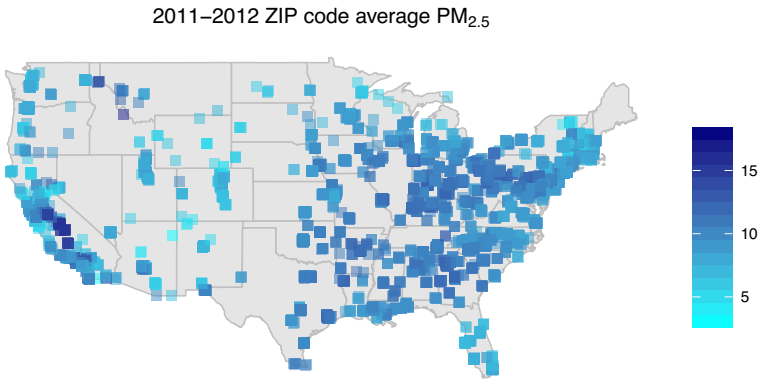


Figure 1.1: Values of $PM_{2.5}$ calculated as the average of the 2011, 2012 values at monitoring sites within 6 miles of a zip code's centroid, for all zip codes in the continental US that were linked to at least one $PM_{2.5}$ monitor.

of health care. Section 1.8.1 includes details regarding the assembling and linkage of the spatially misaligned data, and covariates' description, source, and descriptive statistics. To motivate our methodological development, we now illustrate that –in this data set– different sets of covariates are imbalanced when we restrict the analysis at low exposure levels ($3-8\mu g/m^3$; 816 observations) versus when we restrict the analysis at high exposure levels ($12-13\mu g/m^3$; 324 observations). For each of these two groups of zip codes we introduce binary treatments (T_{i1} and T_{i2} accordingly) defined as follows:

1. $T_{i1} = 0$ if $3 < x_i \leq 7$
2. $T_{i1} = 1$ if $7 < x_i \leq 8$
3. $T_{i2} = 0$ if $12 < x_i \leq 12.5$
4. $T_{i2} = 1$ if $12.5 < x_i \leq 13$.

Within each of the two exposure levels, and for each covariate C_j, we calculate the absolute standardized difference of means (ASDM) based on the binary treatments T_1, T_2. Figure 1.2 shows ASDM when comparing group 2 with group 1 and when comparing group 4 with group 3, separately. Visual inspection of Figure 1.2 indicates that different variables are imbalanced at the low versus the high exposure levels. For example, median house value (in logarithm – `House Value`) is highly imbalanced when considering

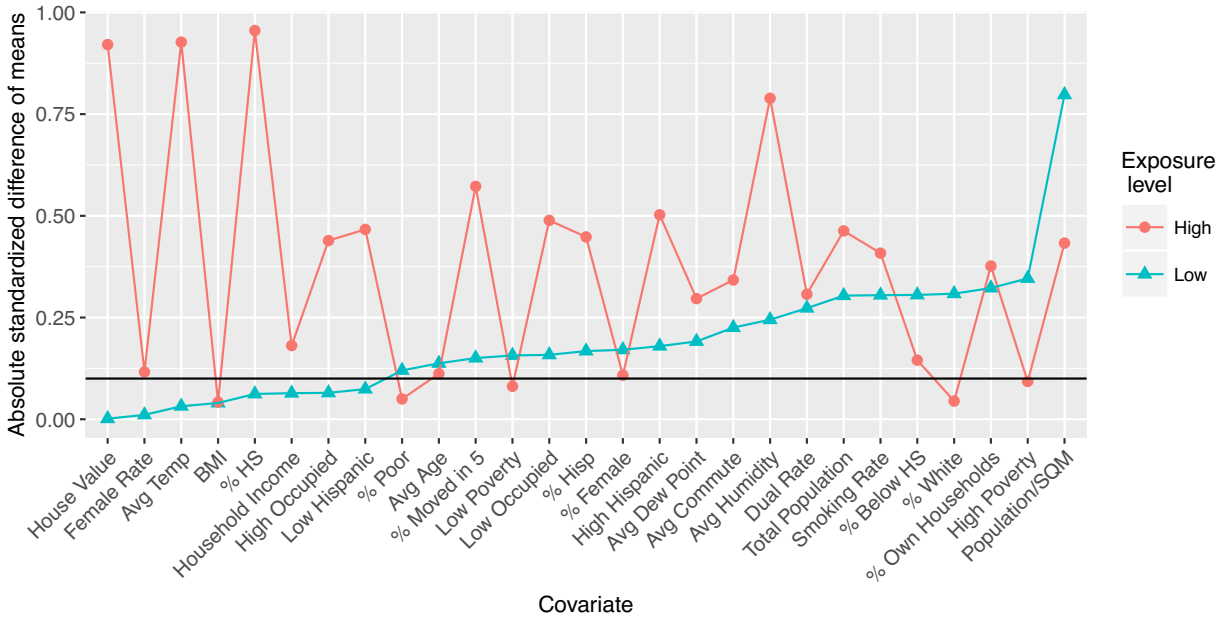


Figure 1.2: Absolute standardized difference of means of binary treatment defined within the low ($3 - 8\mu g/m^3$) and high ($12 - 13\mu g/m^3$) exposure levels. Covariates are ordered in increasing ASDM values for the experiment at low levels.

zip codes at higher exposure values, whereas is not when considering zip codes at lower levels. The opposite is true for other variables such as the proportion of population that is white (`% White`), or has less than high school education (`% Below HS`).

1.3 Notation and Assumptions

We follow the potential outcome framework introduced by Neyman (1923), formalized by Rubin (1974), and extended by Hirano and Imbens (2004) to accommodate continuous exposures. Let $\mathcal{X}$ be the set of possible exposure values. Under SUTVA (Rubin (1980); no interference, no hidden versions of the treatment) let $Y_i(x)$ denote the potential outcome for observation i at exposure $x \in \mathcal{X}$. Then the set $\{Y_i(x), x \in \mathcal{X}\}$ represents the individual ER curve, and $\{\overline{Y}(x) = E[Y_i(x)], x \in \mathcal{X}\}$ the population average ER curve. Assuming sufficient smoothness of $\overline{Y}(x)$ as a function of x, define the instantaneous causal effect

$$\Delta(x) = \lim_{h \to 0} \frac{\overline{Y}(x+h) - \overline{Y}(x)}{h}.$$

$\Delta(x)$ describes the *presence of an effect* of the exposure on the outcome since $\Delta(x) \neq 0$ implies that variation in the exposure in a neighborhood of x has an effect on the expected outcome. Based on $\Delta(x)$ other causal quantities can be defined, such as the effect of an exposure shift from x to $x+\delta$, $CE_\delta(x) = \overline{Y}(x+\delta) - \overline{Y}(x) = \int_x^{x+\delta} \Delta(t)\mathrm{d}t$.

The observed values of the outcome, exposure, and p measured covariates for observation i are denoted as Y_i, X_i, and $\boldsymbol{C}_i = (C_{i1}, C_{i2}, \ldots, C_{ip})$ accordingly. Then $Y_i = Y_i(X_i)$, that is the observed outcome is equal to the potential outcome under the observed exposure. Assuming a random numbering of observations, the subscript i is dropped hereafter.

1.3.1 Experiment configuration, global and local ignorability assumption

The weak ignorability assumption for a continuous exposure (Hirano and Imbens, 2004) states that the treatment is as if randomized conditional on observed covariates

$$X \amalg Y(x) | \mathbf{C},\ x \in \mathcal{X}, \tag{1.1}$$

and every subject in the population can experience any $x \in \mathcal{X}$. Consider a minimal confounding adjustment set $\mathbf{C}^* \subseteq \mathbf{C}$ such that $X \amalg Y(x) | \mathbf{C}^*,\ x \in \mathcal{X}$. Such sets have been previously discussed in the literature (Luna et al., 2011; Wang et al., 2012; Vansteelandt et al., 2012), and they are such that the independence assumption in (1.1) does not hold for any strict subset of $\boldsymbol{C}^*$.

However, the minimal sufficient adjustment set $\mathbf{C}^*$ might vary across exposure levels if different variables confound the exposure-response relationship at different levels of the exposure. We formalize this by introducing the *experiment configuration*. Let K denote a fixed positive integer, $\min = \min x_i$ and $\max = \max x_i$ the minimum and maximum values of the observed exposure, and $\bar{\mathbf{s}} = (s_0 = \min, s_1, s_2, \ldots, s_K, s_{K+1} = \max)$ a known partition of the exposure range in $K+1$ experiments $g_k = [s_{k-1}, s_k), k = 1, 2, \ldots, K+1$. In Figure 1.3, a hypothetical exposure response function is plotted where $\bar{\mathbf{s}}$ defines a total of 4 experiments ($K = 3$). Then, a minimal sufficient adjustment set $\mathbf{C}^*$ can be written as $\mathbf{C}^* = \cup_{k=1}^{K+1} \mathbf{C}_k^*$, where $\mathbf{C}_k^*$ is a minimal sufficient adjustment set for treatment assignment

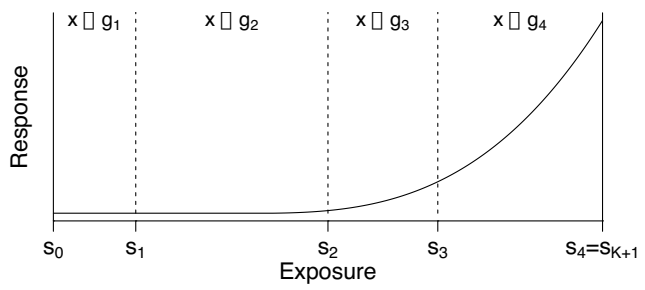


Figure 1.3: Hypothetical ER curve. The exposure range is partitioned by $\bar{s}$ in 4 experiments.

in experiment k, and therefore satisfies

$$X \amalg Y(x) | \mathbf{C}_k^*, \ x \in g_k. \tag{1.2}$$

The sets $\boldsymbol{C}_k^*$ can be overlapping (or even identical) if the same variable is necessary for confounding adjustment at more than one experiment. Note that if (1.2) is satisfied, then (1.1) is also satisfied.

Given $\bar{s}$ and assuming that sets $\boldsymbol{C}_k^*$ satisfying (1.2) exist, model choice can be performed locally within each experiment g_k. Thus, local model selection allows for the identification and adjustment for a different set of confounders at different exposure levels.

1.4 ER estimation in the presence of local confounding

LERCA (Local Exposure Response Confounding Adjustment) is presented for a fixed and unknown experiment configuration $\bar{\mathrm{s}}$ in Section 1.4.1 and Section 1.4.2 respectively. In Section 1.4.3 we describe the MCMC scheme to sample from the posterior distribution of all parameters, for which convergence diagnostics and posterior inference are discussed in Section 1.4.4. Finally, in Section 1.4.5, we discuss the use of WAIC (Watanabe, 2010; Gelman et al., 2014) for choosing the value of K.

1.4.1 Known experiment configuration

Locally, that is for $X_i \in g_k$, $k = 1, 2, \ldots, K+1$, we assume the following pair of exposure and outcome models:

$$\begin{aligned} p(x|\boldsymbol{C} = \boldsymbol{c}, x \in g_k) \;&\propto\; \phi\big(x;\; \delta^X_{k0} + \textstyle\sum_{j=1}^{p} \alpha^X_{kj}\delta^X_{kj}c_j,\; \sigma^2_{k,X}\big) \\ p(y|X = x, \boldsymbol{C} = \boldsymbol{c}, x \in g_k) \;&=\; \phi\big(y;\; \delta^Y_{k0} + \beta_k(x - s_{k-1}) + \textstyle\sum_{j=1}^{p} \alpha^Y_{kj}\delta^Y_{kj}c_j,\; \sigma^2_{k,Y}\big) \end{aligned} \tag{1.3}$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2, $\alpha^X_{kj} = 1$ indicates that covariate C_j is included into the exposure model of the k^{th} experiment, and $\alpha^X_{kj} = 0$ is not. The parameter α^Y_{kj} has the same interpretation, but for the outcome model. The parameter β_k denotes the instantaneous change in the expected outcome associated with a local variation in exposure for $x \in g_k$. Model (1.3) allows for a different set of variables and variables' coefficients for different experiments.[1] If the minimal confounding adjustment set for experiment k is included in the outcome model and the mean functional form is correctly specified, β_k is an unbiased estimator of the instantaneous effect $\Delta(x)$, for $x \in g_k$.

Below we discuss how the prior distributions on all parameters are chosen to target confounding adjustment and continuous ER estimation. More details on the prior specifications can be found in Section 1.8.2.

Prior distribution on inclusion indicators

We build upon the work by Wang et al. (2012, 2015) and Antonelli et al. (2017b) to assign an informative prior on $(\alpha^X_{kj}, \alpha^Y_{kj})$. This prior choice ensures that model averaging assigns high posterior weights to outcome models including a minimal confounding adjustment set, and specifies

$$\frac{P(\alpha^Y_{kj} = 1|\alpha^X_{kj} = 1)}{P(\alpha^Y_{kj} = 0|\alpha^X_{kj} = 1)} = \omega \text{ where } \omega > 1,\; \text{iid } \forall\; j, k. \tag{1.4}$$

By specifying (1.4), a variable C_j is assigned high prior probability to be included into the outcome model of experiment k if it is also included in the exposure model of the

[1] Note that the coefficients and variance terms depend on the inclusion indicators of the corresponding model. For notational simplicity, we do not explicitly state this dependence.

same experiment ($x_i \in g_k$ & $\alpha_{kj}^X = 1$). Wang et al. (2012) and Antonelli et al. (2017b) show that this informative prior leads to outcome models that include the minimal set of true confounders with higher posterior weights than model selection approaches that are based solely on the outcome model. In our context, this experiment-specific prior specification ensures that, locally, covariates in the minimal set $\mathbf{C}_k^*$ are included in the outcome model of experiment k with high posterior probability.

Prior distribution on outcome model intercepts and exposure coefficients for ER continuity

If no structure is assumed on the model (1.3) across experiments, continuity of the estimated ER function at the points of the experiment configuration s_k is not guaranteed. However, in the estimation of the causal effect of exposure to $PM_{2.5}$ on hospitalization outcomes it is expected that the ER function is continuous throughout the exposure range. If the covariates C_j are centered to have mean 0, continuity of the estimated ER function is ensured by assuming a point-mass recursive prior on the intercepts $\delta_{k0}^Y, k \geq 2$ for the outcome model. That is,

$$\lim_{x \to s_k^+} E[Y|X = x] = \lim_{x \to s_k^-} E[Y|X = x] \iff \delta_{k0}^Y = \delta_{(k-1)0}^Y + \beta_{k-1}(s_k - s_{k-1}). \tag{1.5}$$

In other words, the outcome model intercept of experiment $k \geq 2$ is a deterministic function of the outcome model intercept of the first experiment δ_{10}^Y, and slopes $\beta_1, \beta_2, \ldots, \beta_{k-1}$. These parameters are assigned independent non-informative normal prior distributions.

Prior distributions of the remaining coefficients

Prior distributions on the remaining regression coefficients (exposure model coefficients, outcome model covariates' coefficients) and variance terms are chosen such that they lead to known forms of the full conditional posterior distributions to alleviate sampling, as discussed closer in Section 1.4.3. We assume independent non-informative Inverse Gamma prior distributions on $\sigma_{k,X}^2, \sigma_{k,Y}^2$. Non-informative normal prior is chosen for the exposure model intercept δ_{k0}^X. Conditional on the inclusion indicators, the prior on the regression coefficient δ_{kj}^Y is a point mass at 0, or a non-informative normal distribution

when α_{kj}^{Y} is equal to 0 or 1 accordingly. Similarly for the exposure model covariates' coefficients δ_{kj}^{X}.

1.4.2 Unknown experiment configuration

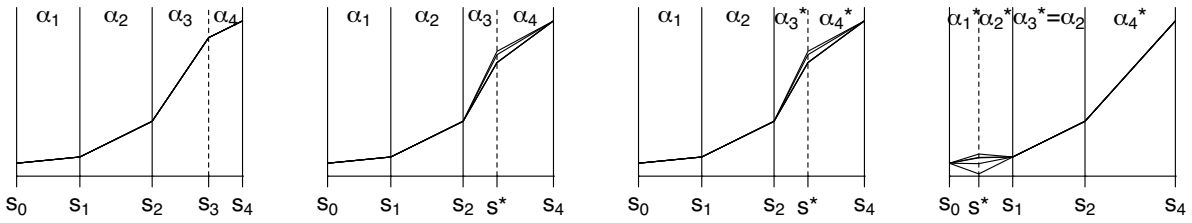
For a fixed experiment configuration $\bar{\mathrm{s}}$, each experiment is treated separately in terms of confounder selection and strength of the confounding adjustment. However, the configuration itself is a key component of the fitted exposure response curve, and fixing it a priori could lead to bias and uncertainty underestimation.

LERCA is extended to allow for unknown experiment configuration $\bar{\mathrm{s}}$, while carrying all the merits described above. The locations of the experiment configuration $\boldsymbol{s} = (s_1, s_2, \ldots, s_K)$ are, a priori, assumed to be distributed as the even-numbered order statistics of $2K + 1$ samples from a uniform distribution on the interval (s_0, s_{K+1}). This prior choice of $\boldsymbol{s}$ discourages specifications of $\boldsymbol{s}$ that include values that are close to each other (Green, 1995). The prior is augmented by indicators that consecutive points s_k, s_{k+1} cannot be closer than some distance d_k. Conditional on $\boldsymbol{s}$, we follow the model specification and prior distributions described in Section 1.4.1.

1.4.3 MCMC scheme and computational challenges

The factorization of the full data likelihood over experiments and exposure/outcome models and the choice of the prior distributions lead to full conditional posterior distributions of coefficients $\delta_{k0}^{X}, \delta_{kj}^{X}, \delta_{kj}^{Y}$, and variance terms $\sigma_{k,X}^{2}, \sigma_{k,Y}^{2}$ of known forms. The variance terms and exposure model intercepts have inverse Gamma and normal full conditional posterior distributions accordingly, whereas the distributions of $\delta_{kj}^{X}, \delta_{kj}^{Y}$ are either point mass at 0 or normal, based on whether the corresponding α is 0 or 1.

Since δ_{k0}^{Y}, $k \geq 2$ is a deterministic function of δ_{10}, $\beta_1, \beta_2, \ldots, \beta_{k-1}$, and the points $s_0, s_1, \ldots, s_k$, the full conditional posterior distribution of δ_{10} depends on data across all experiments, and that of β_k on data from experiment k and onwards. Since the data likelihood in all experiments is normal and we have assumed normal prior distributions, the full conditional posterior distributions are also normal. After each update, intercepts $\delta_{k0}^{Y}, k \geq 2$ need to be updated from (1.5) to ensure ER continuity.



(a) Current state (b) Separate (c) Jump within (d) Jump over

Figure 1.4: Proposed state for the separate, jump within and jump over moves are depicted schematically for a hypothetical experiment configuration with $K = 3$. In all proposed states, new slopes are proposed to ensure continuity of the ER. (a) The current state of the MCMC. s_3 is chosen to be updated. (b) Separate: A new point s^* is proposed within (s_2, s_4) with the corresponding α parameters constant. (c) Jump within: Simultaneous move of the experiment configuration and the corresponding α's within (s_2, s_4). (d) Jump Over: The proposed point s^* is located outside the interval (s_2, s_4) and new α's are proposed for the experiment that was split (s_0, s_1), and the experiments that were combined (s_2, s_4).

In order to avoid the need of proposing values for the covariates' coefficients and variance terms in the update of the experiment configuration, these parameters are integrated out from the data likelihood and the Bayes factors are approximated using the BIC (Raftery, 1995). Updates of the experiment configuration and inclusion indicators are performed most of the times using a "separate", and sometimes using a "jump over" or "jump within" move, depicted in Figure 1.4. Note here that an update in $\boldsymbol{s}$ needs to be accompanied with an update of the coefficients δ^Y_{k0} and β_k to ensure ER continuity.

When the experiment configuration and inclusion indicators are updated separately, the update of $\boldsymbol{s}$ is performed using Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) where a point s_k is proposed to be moved to $s^* \in (s_{k-1}, s_{k+1})$. New values for β_k and β_{k+1} are also proposed to ensure ER continuity, as shown in Figure 1.4(b). After acceptance or rejection of this move, the inclusion indicators are updated from

$$p(\alpha^Y_{kj} = \alpha|\text{Data}, A^*, \alpha^X_{kj}) \;\propto\; \frac{p(\delta^Y_{kj} = 0|\alpha^Y_{kj} = \alpha)p(\alpha^Y_{kj} = \alpha|\alpha^X_{kj})}{p(\delta^Y_{kj} = 0|\alpha^Y_{kj} = \alpha, \text{Data}, A^*)},\ \alpha \in \{0, 1\}, \tag{1.6}$$

where A^* are all parameters but α^X_{kj}, α^Y_{kj} and δ^Y_{kj}.

In order to improve mixing of the MCMC, $\boldsymbol{s}$, and $\underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y$ are sometimes updated simultaneously implementing a "jump within" or "jump over" Metropolis-Hastings move with which the proposal maintains or not the order of the experiment configuration. The as-

signment of proposed values for the inclusion indicators is probabilistic based on their current values, encouraging the inclusion of a covariate in the proposed state to resemble that of the current state. For example, in the "jump within" move depicted in Figure 1.4(c), α_3^*, α_4^* should resemble α_3, α_4 since they refer to similar exposure ranges. At the same time, coefficients β_k are proposed such that they lead to unaltered likelihood of the unaffected experiments. Figure 1.4(c) depicts random draws for proposed ER states. Section 1.8.3 includes an in-depth description of the MCMC scheme.

1.4.4 Posterior inference and MCMC convergence

Due to the update of the experiment configuration, commonly used convergence diagnostics such as trace plots are not appropriate since parameters (e.g., β_k) may correspond to a different range of exposure values at different iterations. Therefore, convergence must be examined in the context of quantities that are detached from the experiment configuration.

Posterior inference and estimation of causal quantities of interest is performed over a set of potential exposure values $\mathcal{G} \subset \mathcal{X}$. For every $x \in \mathcal{G}$, a posterior sample of the mean response at exposure x is equal to $\delta^Y_{k_x 0} + \beta_{k_x}(x - s_{k_x - 1})$, where k_x is the experiment in which x belongs to at the specific iteration. Based on the posterior samples of the mean ER, the potential scale reduction factor (PSR; Gelman and Rubin (1992)) is calculated, and MCMC convergence is evaluated based on $|\text{PSR} - 1| < c$ for all $x \in \mathcal{G}$. More details about MCMC diagnostics can be found in Section 1.8.3.

1.4.5 Choosing the number of points in the experiment configuration

LERCA requires the specification of the number of points K in the experiment configuration. Since the number of parameters grows with K, possible values for K could be bounded by considering the maximum number of coefficients we are willing to entertain. Cross validation methods are commonly used in order to choose values of key tuning parameters, but are often infeasible in the Bayesian framework due to time and computational resources constraints. In a comprehensive review, Gelman et al. (2014) discusses various methods of estimating the expected out of sample prediction error for Bayesian

methods.

We use the widely-applicable information criterion (WAIC; Watanabe (2010)) to acquire an estimate of the out-of-sample prediction error. LERCA is fit for different values of K, and K is chosen as the value that minimizes the WAIC. The WAIC combines the log point-wise posterior predictive density ($lppd$) and a penalty term for over-fitting p_{WAIC}, $WAIC = -2\left(lppd - p_{WAIC}\right)$, where

$$lppd = \sum_{i=1}^{n} \log E_{post} p(x_i, y_i|\theta)$$

and

$$p_{WAIC} = \sum_{i=1}^{n} \text{var}_{post}\left(\log p(x_i, y_i|\theta)\right),$$

for $\theta = (\boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y)$ and E_{post}, var_{post} denoting expectation and variance over the posterior distribution. All expectations can be estimated using the posterior samples of one MCMC run. For example,

$$\widehat{lppd} = \frac{1}{T} \sum_{i=1}^{n} \sum_{t=1}^{T} \log p(x_i, y_i|\theta^{(t)}),$$

where $\theta^{(t)}$ is the value of the parameters at iteration t.

1.5 LERCA illustration and performance evaluation in the presence of local confounding

1.5.1 Data generation

Data generation in the presence of local confounding is complicated and is described in detail in Appendix 1.8.4. Here, we present a simulation scenario where: (a) local confounding is present, and (b) the true shape of the ER is quadratic. We assume that exposure values x_i range from 0 to 10, and the true experiment configuration is $\bar{\boldsymbol{s}} = (0, 2, 4, 7, 10)$. Table 1.1 summarizes which of the 8 potential confounders are predictive of the exposure and/or the outcome within each experiment (correlations and regression coefficients are summarized in Table 1.3). The adjusted R-squared of the true exposure and outcome models within each experiment varied between 0.64 and 0.88. We simulate 400 data sets of 800 observations each.

1.5.2 Goal of the simulations

We illustrate that commonly-used approaches for ER estimation are not appropriate for confounding adjustment in the presence of local confounding. For the methods utilizing the generalized propensity score (gps), a model of exposure linear in all the covariates is adopted. The approaches considered are:

1. Generalized Additive Model (`GAM`): Regressing the outcome Y on the exposure X and all potential confounders with 4 degrees of freedom for every covariate.

2. Spline Model (`SPLINE`): Additive spline estimator described in Bia et al. (2014). The dose response function is estimated as splines of the exposure and the gps.

3. The Hirano and Imbens estimator (Hirano and Imbens, 2004) (`HI-GPS`): ER estimation is based on an outcome model regression including quadratic terms for the exposure and the gps, and their interaction.

4. Inverse Probability Weighting estimator (`IPW`): The generalized propensity score is used to weigh observations in an outcome regression model that includes linear and quadratic terms of exposure.

The R packages `gam` and `causaldrf` were used (Hastie, 2017; Schafer, 2015).
Additionally, we compare the root mean squared error (rMSE) of all methods in estimating the mean exposure response curve $\overline{Y}(x)$. Finally, was also assess whether LERCA can

Table 1.1: Representation of which covariates are predictive of the exposure and / or the outcome within each experiment (denoted by a ✓). Covariates with ✓in both models within the same experiment are local confounders.

Experiment	Model	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1	$X\|\boldsymbol{C}$	✓	✓	✓					
	$Y\|X,\boldsymbol{C}$	✓	✓	✓					
2	$X\|\boldsymbol{C}$	✓	✓		✓				
	$Y\|X,\boldsymbol{C}$		✓	✓	✓				
3	$X\|\boldsymbol{C}$	✓		✓		✓			
	$Y\|X,\boldsymbol{C}$		✓	✓		✓			
4	$X\|\boldsymbol{C}$		✓			✓	✓		
	$Y\|X,\boldsymbol{C}$		✓	✓			✓		

recover the correct experiment configuration, identify the true confounders within each experiment, and choose the true value for K.

1.5.3 Simulation Results

For every simulated data set, LERCA was fit for $K \in \{2, 3, 4\}$, for which the ER was estimated over an equally spaced grid over the interval $(0, 10)$ denoted by $\mathcal{G}$. Results are presented for the simulated data sets for which the MCMC converged for all choices of K (for convergence diagnostics, see Section 1.4.4; $c = 0.05$).

Figure 1.5 shows LERCA simulation results including the estimated ER, experiment configuration, and posterior inclusion probabilities of covariates C_1, C_4 in the outcome model as a function of exposure $x \in (0, 10)$. Figure 1.6 shows the estimated ER curve for the four

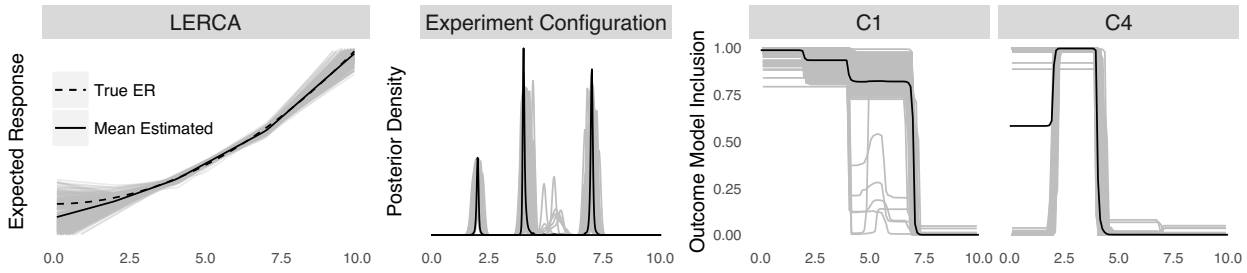


Figure 1.5: LERCA results. (Left) Mean ER estimates. (Center) Posterior density distribution of the experiment configuration $\boldsymbol{s}$. (Right) Outcome model posterior inclusion probability of C_1 and C_4. Gray lines correspond to results per simulated data set, and black solid lines correspond to summaries across data sets.

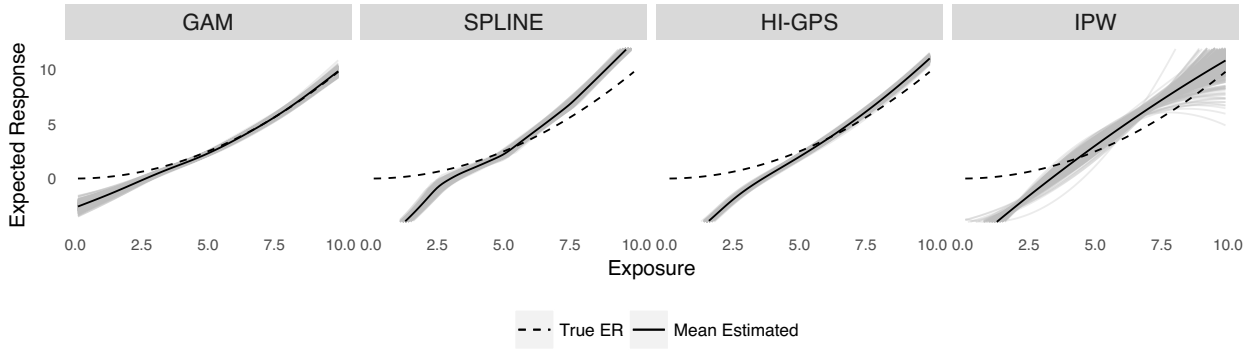


Figure 1.6: The true mean ER function (dashed line), posterior mean ER functions from each simulated data set (gray), and the mean of the estimated ER functions (solid lines) using all alternative methods.

alternative methods described above. Grey lines correspond to results from individual data sets, whereas black solid lines correspond to averages across simulated data sets. LERCA discovers the correct shape of the exposure-response function as depicted in Figure 1.5, even though the true ER is quadratic and LERCA is formulated as piece-wise linear. The alternative methods return biased results across the exposure range (as shown in Figure 1.6), indicating that they are not appropriate for ER estimation in the presence of local confounding. In fact, the root MSE of LERCA was consistently lower than the alternative methods at low exposure levels (Figure 1.11), and across exposure values in $\mathcal{G}$ it ranged from 0.1–1.24 for LERCA, followed by GAM at 0.13–2.5.

Moreover, using WAIC to choose the value of K led to choosing the correct value of $K = 3$ 40% of the times, and $K = 2$ 58% of the times indicating that WAIC tends to heavily penalize large values of K. However, the correct points of the experiment configuration $\boldsymbol{s} = \{2, 4, 7\}$ are identified and are located at the modes of the posterior distribution as shown in Figure 1.5. By examining the posterior inclusion probabilities of C_1, C_4, we observe that instrumental variables (e.g., C_1 in experiments 2 and 3) are often included in the outcome model. However, LERCA includes the minimal confounding set within each experiment with very high probability. On average (across the points in $\mathcal{G}$ and simulated data sets) the minimal confounding set was included in the adjustment set 99% of the times (ranging from 89-100% across simulated data sets), indicating that the variables necessary for confounding adjustment are almost always included in the adjustment set. Lastly, the point-wise 95% and 50% credible intervals cover the true mean ER values 84% and 39% of the times accordingly. The under-coverage is largely due to the underestimation of K.

1.5.4 Simulation results in the absence of local confounding

The previous data simulation scenario compared the performance of LERCA in the presence of local confounding. In Section 1.8.5, methods' performance was compared under global confounding (same confounders across exposure levels) and a quadratic ER representing scenarios that current methods are developed to address. LERCA with $K = 3$ (fixed) performed similarly in terms of root MSE compared to GAM, and better than all

other alternative methods. These results indicate that LERCA offers a protection against bias arising from local confounding, while not sacrificing much when local confounding is not present.

1.6 Data Application

We applied LERCA to estimate the ER curve between exposure to $PM_{2.5}$ during the years 2011-2012 and log cardiovascular hospitalization rates at the zip code level in 2013, as discussed in Section 1.2. The full set of zip code level covariates are described in Table 1.2. LERCA was fit for $K \in \{2, 3, \ldots, 6\}$ and results presented correspond to $K = 3$ which returned the lowest WAIC.

Figure 1.7 shows (a) mean ER estimates and 95% credible intervals for exposure to $PM_{2.5}$ and log cardiovascular hospitalization rates, (b) mean and 95% credible interval for the coefficient β_k, the estimator of the instantaneous effect for which a 95% credible interval including only positive values implies that a local increase in $PM_{2.5}$ exposure would lead to a significant increase in hospitalization rates, (c) the posterior distribution of the experiment configuration, and (d) the observed distribution of $PM_{2.5}$.

An overall increasing trend in the ER is observed, and for $PM_{2.5} < 9.9\mu g/m^3$, an increase in $PM_{2.5}$ exposure leads to a significant increase in log hospitalization rates, as indicated by the credible intervals on $\hat{\Delta}(x)$. For values of $PM_{2.5} \geq 9.9\mu g/m^3$, 95% posterior credible intervals of the $\hat{\Delta}(x)$ cover 0. These results are in accordance with the scientific belief that the effect of exposure to $PM_{2.5}$ is smaller at higher exposure levels. Lastly, the posterior distribution of $\boldsymbol{s}$, shows that observations below $8\mu g/m^3$ and over $11.5\mu g/m^3$ are always grouped together. This could be due to limited sample size at the extreme exposures, as depicted by the bottom panel of Figure 1.7.

As done in the simulation study, the covariates' posterior inclusion probability can be plotted as a function of the exposure. Figure 1.8 shows the posterior exposure and outcome model inclusion probability for median house value, Medicare female rate and population density, which presented differential imbalances at low and high exposure levels in Section 1.2. We see that the posterior inclusion probabilities for these variables are in

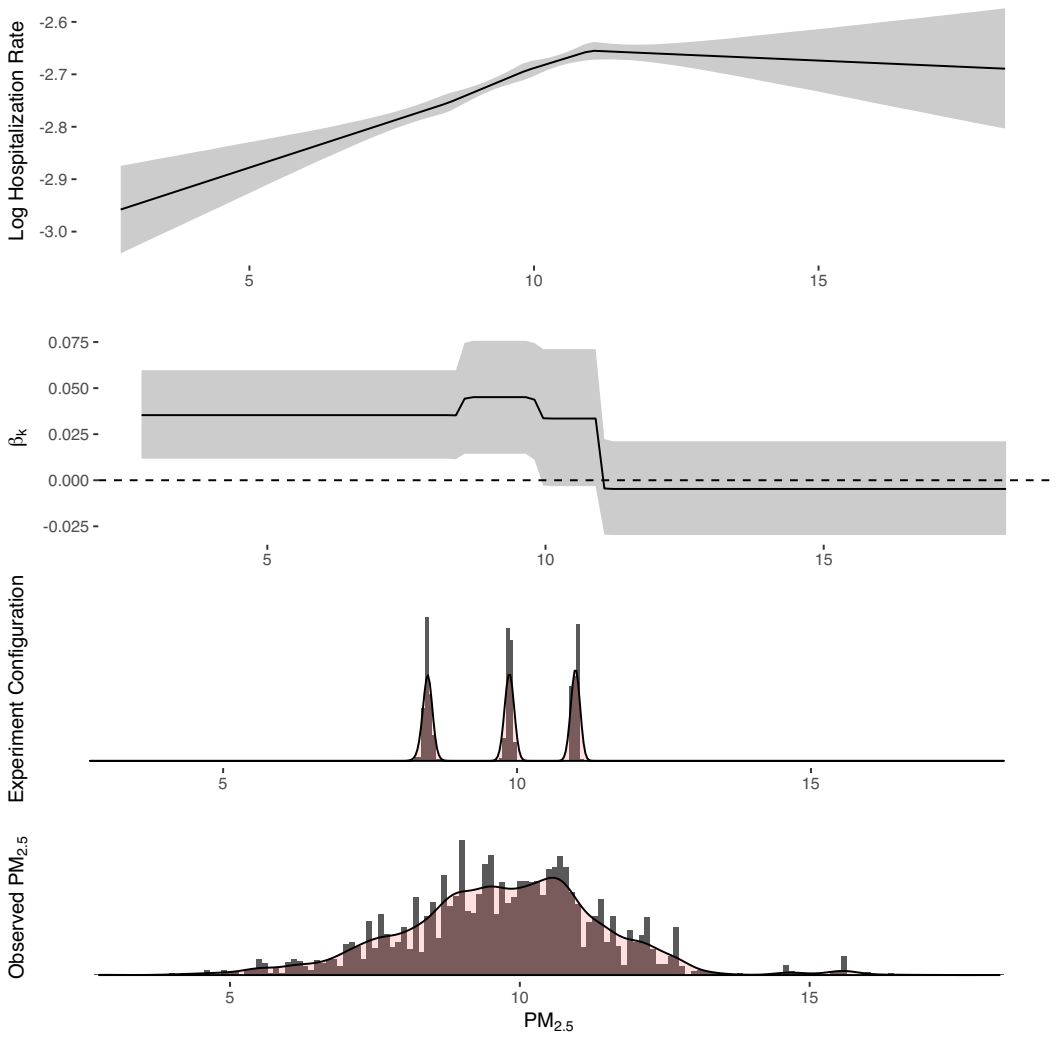


Figure 1.7: From top to bottom: Mean ER curve of $PM_{2.5}$ exposure (x-axis) on log all-cause cardiovascular hospitalizations (y-axis) –solid line– with 95% pointwise credible intervals. The posterior mean and 95% credible interval of the β coefficient as a function of exposure. The posterior distribution for $\boldsymbol{s}$. Observed $PM_{2.5}$ values in the data.

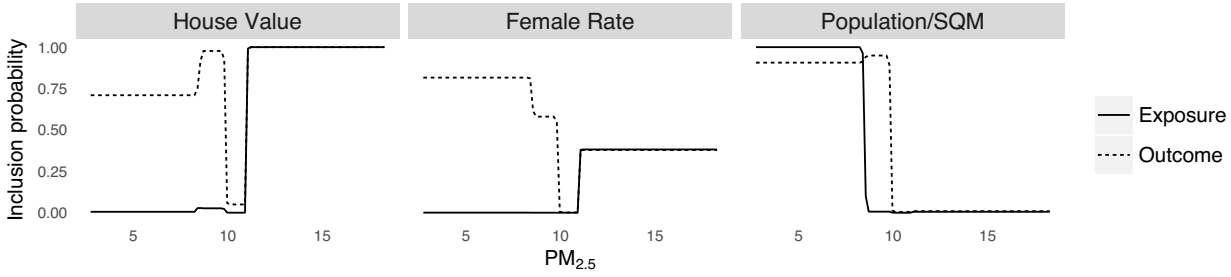


Figure 1.8: Posterior inclusion probability of zip code median house value, Medicare female rate, and population density in the exposure and outcome model at different exposure levels.

concordance to the explanatory analysis presented in Figure 1.2. For example, Figure 1.2 and Figure 1.8 agree that the median house value is a predictor of exposure for high exposure values, while it's not predictive of exposure at low levels.
However, the posterior inclusion probability of several variables does not agree with the ASDM of Figure 1.2. This could happen because 1) the ASDM is calculated based on a categorized treatment, and 2) the inclusion probabilities represent a more localized importance measure than ASDM.

1.7 Discussion

We have introduced an innovative Bayesian approach for flexible estimation of the ER curve in observational studies that has the following important features: 1) let the data inform the experiment configuration; and given the experiment configuration 2) allows for the possibility (which is a reality in our data example, see Section 1.2) that different sets of covariates are indeed confounders at different exposure levels; 3) allows for varying confounding effect across levels of the exposure; 4) performs covariate selection, locally, that it, within each exposure range to increase efficiency, especially at low exposure levels; 5) propagates model uncertainty for the experiment configuration and covariate selection in the posterior inference on the whole ER curve; 6) reduces sensitivity related to the choice of the shape of the ER curve; 7) provides important scientific guidance related to which covariates are confounders at different exposure levels; and finally, 8) allows for the estimation of a potentially flat ER function at the very low levels of exposure, thus allowing for the identification of a threshold.
However, the proposed approach also has some draw backs. First, within each experiment, we assume linearity for both the outcome and the exposure model. If this relationship is specified incorrectly, unbiasedness of the results is not guaranteed. Even though linearity of the exposure in the outcome model could be easily relaxed by using higher order splines, we considered it of high importance to accommodate the "absence of an effect" scenario at low exposure levels. Furthermore, if linearity is strongly not supported from the data, we expect that the experiment configuration $\boldsymbol{s}$ will adapt to accommodate

non-linearity.

Second, the informative prior on the inclusion indicators could lead to the inclusion of instrumental variables in the outcome model with high posterior probability, which will not lead to bias, but will decrease the efficiency of our estimators. However, in the study of air pollution, strong instrumental variables are not expected to be present.

Dependence of model parameters across different experiments was limited to outcome model intercepts to ensure continuity of the estimated ER. However, imposing additional structure could be easily incorporated. For example, it might be reasonable to assume that a variable is more likely to be included in the exposure model of an experiment if it is also included in the exposure model of a neighboring experiment. Such structure could be incorporated by reformulating the prior on the inclusion indicators (1.4) to allow

$$\frac{P(\alpha_{kj}^X = 1 | \alpha_{k-1,j}^X = 1)}{P(\alpha_{kj}^X = 0 | \alpha_{k-1,j}^X = 1)} = \theta > 1, \text{ for } k \geq 2$$

Alternatively, reformulation of the independent prior distributions of β_k could incorporate cross-experiment dependence by specifying $\beta_k \sim N(\beta_{k-1}, \sigma_\beta^2)$, $k \geq 2$. We believe that this is an exciting line of research for future work.

Although non-parametric and varying coefficient approaches (Hastie and Tibshirani, 1993) for ER estimation could, in theory, allow for differential confounding effect across different exposure levels, none of the existing methods for ER estimation explicitly accommodates differential confounding sets for different exposure levels, nor provides guidance for which covariates are confounders of the effect of interest at different levels of the exposure. Furthermore, the use of non-parametric methods to estimate a generalized propensity score or model the outcome of interest could prove unfruitful in situations where most of the available data are over a specific range of the exposure variable, or the number of potential confounders is large, and interest lies in the estimation of causal effects for change in the exposure in the tails of the exposure distribution. In such situations, LERCA provides a way to model the outcome acknowledging that the exposure-response relationship might be confounded by different covariates at different exposure levels.

1.8 Appendix

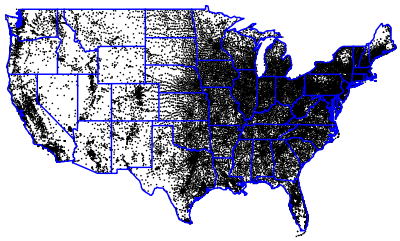
1.8.1 Data details

We constructed counts corresponding to the cardiovascular-specific (CVD) number of hospitalizations for Medicare enrollees aged at least 65 years during 1999-2013 for a total of 42,139 zip codes across the continental US. Hospitalization rates were based on the total number of personal years for Medicare enrollees for a zip code on a given year. CVD hospitalizations were considered on the basis of primary diagnosis according to International Classification of Diseases, Ninth Revision (ICD-9) codes (ICD-9 390 to 459). The analysis was restricted to 2013 and on the continental US leading to 34,897 zip codes with hospitalization information.

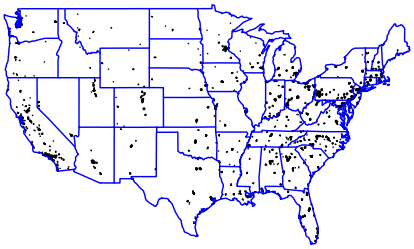
Population demographic information was acquired using the 2000 Census with information on over 400 variables, although a lot of them are highly correlated. We further used linearly extrapolated Census variables for 2013. Census information is provided at a ZCTA level, and we use a crosswalk to map ZCTA to zip code. Weather information including temperature, relative humidity and dew point is acquired from the National Oceanic and Atmospheric Administration (NOAA) Automated Surface Observing System (ASOS), and is linked to zip codes within 150 kilometers.

Lastly, zip code $PM_{2.5}$ exposure is assigned using the US EPA monitoring sites. By EPA recommendations, monitoring sites with less than 67% of scheduled measurements observed are excluded. For every monitor, the average of the 2011-2012 average annual value of $PM_{2.5}$ is calculated, and the monitor is linked to *all* zip codes with centroids within 6 miles. Then, the zip code exposure is set equal to the average over all linked monitors. Since monitoring sites are preferentially located near populated areas or points of interest, many zip codes in remote areas are not linked to any monitor and are therefore dropped from the final data set.

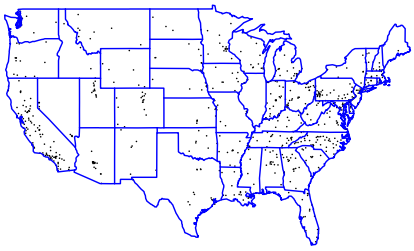
Figure 1.9 shows maps pf zip code centroids before linkage to EPA monitoring sites, as well as maintained zip code centroids after 3 different linkage procedures corresponding to different specifications of the linkage distance, as well as whether a monitor can be linked to more than one zip code. We visualize how linkage can affect the final data set:



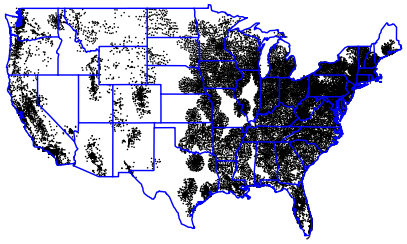
(a) All zip codes in Medicare.



(b) Zip codes with PM monitor within 6 miles. Linkage not unique.



(c) Zip codes with PM monitor within 6 miles. Unique linkage.



(d) Zip codes with PM monitor within 60 miles. Linkage not unique.

Figure 1.9: (a) All zip codes with available Medicare information. (b) Zip codes with available exposure information after performing linkage within 6 miles and monitors are allowed to be linked to more than one zip code. (c) Zip codes with available exposure information after performing linkage within 6 miles where each monitor is only linked to up to one zip code. (d) Zip codes with available exposure information after linkage with monitors within 60 miles and every monitor can be linked to more than one zip code.

- **Distance:** As the distance of allowed linked zip codes and monitors increases, we expect that more zip codes will be linked to at least one monitor. However, the assigned values of $PM_{2.5}$ will be more uncertain in areas where monitors are located at long distances.

- **Number of links:** Allowing a monitor to be linked to multiple zip codes increases the number of zip codes with $PM_{2.5}$ information. However, this can lead to adjacent zip codes with very similar or identical $PM_{2.5}$ measurements.

Table 1.2: Available demographic and weather information

Source	Name	Description	Mean	SD
2000 Census	% White	Percentage of White Population	0.71	0.25

Table 1.2 (Continued.)

	% Hisp	Percentage of Hispanic Population	0.12	0.18
	% HS	Percentage of population that attended high school	0.27	0.10
	% Poor	Percentage of impoverished population	0.14	0.11
	% Female	Percentage of female population	0.51	0.04
	% Moved in 5	Percentage of population that has lived in the area for less than 5 years	0.50	0.12
	Avg Commute	Mean Travel Time to Work	24.22	5.92
	Population/SQM	Population per square mile (logarithm)	7.53	1.52
	Total Population	Total population (logarithm)	9.71	1.12
	Low Occupied	Indicator. "=1" if the percent of occupied population is at most 90%.	0.211	0.408
	High Occupied	Indicator. "=1" if the percent of occupied population is over 95%.	0.416	0.493
	Low Hispanic	Indicator. "=1" if the percent of Hispanic population is at most 0.02%	0.317	0.465
	High Hispanic	Indicator. "=1" if the percent of Hispanic population is over 20%	0.197	0.398
Census Extrapolation	% Below HS	Population percent with less than high school education (above age of 65)	23.24	14.85
	% Own Households	Percentage of occupied housing units in 2013	0.58	0.2
	Low Poverty	Indicator. "=1" if the percent of the population below the poverty line in 2013 is at most 5%	0.196	0.397
	High Poverty	Indicator. "=1" if the percent of the population below the poverty line in 2013 is over 15%	0.244	0.429
Census combination [2]	House Value	Median value of owner occupied housing (USD) (logarithm)	12.65	0.63
	Household Income	Median household income (USD) (logarithm)	11.40	0.42
BRFSS	BMI	Average BMI in 2013	27.65	1.32
	Smoking Rate	Ever smoke rate (2013)	0.45	0.06

[2]The 2000 Census is combined with the 2013 extrapolated values of the same variable by taking the mean of the variable across the two years.

Table 1.2 (Continued.)

Weather	Avg Temp	Average temperature (F)	55.35	7.47
	Avg Dew Point	Average Dew Point (F)	44.09	7.50
	Avg Humidity	Average Relative Humidity (%)	70.41	8.34
Medicare	Avg Age	Average Medicare Age	74.89	1.66
	Female Rate	Percentage of Female Beneficiaries	0.55	0.06
	Dual Rate	Percentage of Dual Eligible Beneficiaries	0.22	0.15

1.8.2 Prior specifications for regression parameters and experiment configuration

Regression coefficients and residual variance

Prior independences of all parameters are expressed in the following representation

$$
\begin{aligned}
&p(\underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\delta}}^X, \beta_k, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y) \\
&= p(\delta^Y_{10}) \prod_{k=1}^{K+1} \left\{ \left[\prod_{j=1}^{p} p(\alpha^X_{kj}, \alpha^Y_{kj})\, p(\delta^X_{kj}|\alpha^X_{kj})\, p(\delta^Y_{kj}|\alpha^Y_{kj}) \right] p(\delta^X_{k0})\, p(\beta_k)\, p(\sigma^2_{k,X})\, p(\sigma^2_{k,Y}) \right\}. \quad (1.7)
\end{aligned}
$$

We assume non-informative normal priors on β_k, $k = 1, 2, \ldots, K+1$, and δ^Y_{10}. The prior distribution on the regression coefficients is a mixture of non-informative normal distribution and point-mass at 0. Non-informative inverse gamma prior distributions are assumed on $\sigma^2_{k,X}, \sigma^2_{k,Y}$. Specifically

- $\beta_k \sim N(\mu_0, \sigma^2_0)$, $\delta_{10} \sim N(\mu_0, \sigma^2_0)$.
- $\delta^X_{kj}|\alpha^X_{kj} \sim \alpha^X_{kj}\, N(\mu_0, \sigma^2_0) + (1 - \alpha^X_{kj})\, \mathbb{1}_0(\delta^X_{kj})$, where $\mathbb{1}_0(\delta^X_{kj})$ is a point-mass distribution at 0. Similarly for $\delta^Y_{kj}|\alpha^Y_{kj}$.
- $\sigma^2_{k,X} \sim IG(a_0, b_0)$, and similarly for $\sigma^2_{k,Y}$.

The hyper-parameters $\mu_0, \sigma^2_0, a_0, b_0$ can be chosen differently for different variables.

Experiment configuration

The prior on the points $\boldsymbol{s} = (s_1, s_2, \ldots, s_K)$ defining the experiment configuration is set as the even ordered statistics of $(2K+1)$ samples from a uniform distribution over the observed exposure range. Compared to a uniform prior distribution on $\boldsymbol{s}$, this choice of

a prior discourages the existence of points s_i, s_j in the experiment configuration that are very close to each other.

Let K and the exposure range (s_0, s_{K+1}) be fixed. Let $Z_i \sim U(s_0, s_{K+1})$, $i = 1, 2, \ldots, 2K+1$ and denote the even ordered statistics as $W_j = Z_{(2j)}, j = 1, 2, \ldots, K$. Then,

$$f_{W_1,W_2,\ldots,W_K}(w_1, w_2, \ldots, w_K) = \\ f_{W_1}(w_1) f_{W2|W1}(w_2|w_1) \ldots f_{W_K|W_1,W_2,\ldots,W_{K-1}}(w_k|w_1, w_2, \ldots, w_{k=1})$$

Since W_1 is th 2^{nd} order statistic of $2K+1$ samples from $U(s_0, s_{K+1})$, we know that

$$\begin{aligned} f_{W_1}(w_1) =& \frac{(2K+1)!}{(2K-1)!} \frac{1}{s_{K+1}-s_0} \frac{w_1 - s_0}{s_{K+1}-s_0} \left(1 - \frac{w_1 - s_0}{s_{K+1}-s_0}\right)^{2K-1} \\ =& \frac{(2K+1)!}{(2K-1)!} (s_{K+1}-s_0)^{-(2K+1)} (w_1 - s_0)(s_{K+1} - w_1)^{2K-1} \end{aligned}$$

Given $W_1 = w_1$, W_2 acts like the second order statistic of $2K-1$ uniform samples from a uniform distribution over (w_1, s_{K+1}). Therefore, we similarly get that

$$f_{W_2|W_1}(w_2|w_1) = \frac{(2K-1)!}{(2K-3)!} (s_{K+1} - w_1)^{-(2K-1)} (w_2 - w_1)(s_{K+1} - w_2)^{2K-3}.$$

Iteratively, we have that

$$f_{W_1,W_2,\ldots,W_K}(w_1, w_2, \ldots, w_K) = \\ (2K+1)!(s_{K+1}-s_0)^{-(2K+1)} (w_1 - s_0)(w_2 - w_1) \ldots (w_K - w_{K-1})(s_{K+1} - w_K)$$

Therefore, the prior distribution on $\boldsymbol{s}$ with minimum distance of consecutive points s_k, s_{k+1} being d_k is defined as

$$f_{\boldsymbol{s}}(s_1, s_2, \ldots, s_K) \propto \prod_{k=0}^{K} (s_{k+1} - s_k) \mathbb{I}(s_{k+1} - s_k > d_k) \tag{1.8}$$

1.8.3 Sampling from the posterior distribution

The parameters included in the model are: $\boldsymbol{s}$ (the exposure values in the experiment configuration), $\underset{\sim}{\boldsymbol{\alpha}}^X$, $\underset{\sim}{\boldsymbol{\alpha}}^Y$ (the vectors of length p including the covariates' inclusion indicators in the exposure and the outcome model for each experiment), $\underset{\sim}{\boldsymbol{\beta}} = \{\beta_k\}_{k=1}^{K+1}$ (coefficients of exposure in the outcome model), $\underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y$ (intercepts and coefficients of the covariates in the exposure and outcome model of each experiment), $\underset{\sim}{\boldsymbol{\sigma}}^2_X = \{\sigma^2_{k,X}\}_{k=1}^{K+1}, \underset{\sim}{\boldsymbol{\sigma}}^2_Y = \{\sigma^2_{k,Y}\}_{k=1}^{K+1}$ (residual variance of the exposure and outcome within each experiment).

Likelihood factorization

We start by noting that the full conditional data likelihood factorizes to components for different experiments and the exposure and outcome models. If $\boldsymbol{Y}, \boldsymbol{X}$ denote the vectors of outcomes and exposures for all units in the sample, and $\boldsymbol{Y}^k, \boldsymbol{X}^k$ denote the vectors of outcomes and exposures in experiment k, then

$$\begin{aligned} P(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y, \mathbf{C}) = \\ \prod_{k=1}^{K+1} \prod_{i \in g_k} p_k(Y_i | X_i, \underset{\sim}{\boldsymbol{\alpha}}^Y_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \beta_k, \sigma^2_{k,Y}, \boldsymbol{C}_i) p_k(X_i | \underset{\sim}{\boldsymbol{\alpha}}^X_k, \underset{\sim}{\boldsymbol{\delta}}^X_k, \sigma^2_{k,X}, \boldsymbol{C}_i) = \\ \prod_{k=1}^{K+1} \left[p_k(\boldsymbol{Y}^k | \boldsymbol{X}^k, \underset{\sim}{\boldsymbol{\alpha}}^Y_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \beta_k, \sigma^2_{k,Y}, \boldsymbol{C}^k) p_k(\boldsymbol{X}^k | \underset{\sim}{\boldsymbol{\alpha}}^X_k, \underset{\sim}{\boldsymbol{\delta}}^X_k, \sigma^2_{k,X}, \boldsymbol{C}^k) \right], \end{aligned} \tag{1.9}$$

where we denote $p_k(\cdot_1 | \cdot_2)$ as the density of $\cdot_1$ conditional on $\cdot_2$ in experiment k and $\underset{\sim}{\boldsymbol{\delta}}^Y_k$ includes the intercept δ^Y_{k0}.

Next, we note that if we consider the marginal likelihood integrating out 1) exposure model regression coefficients including the intercept, 2) outcome model covariates' regression coefficients, and 3) all variance terms, then the likelihood still factorizes in a similar manner. In fact

$$p(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \boldsymbol{C}) = \prod_{k=1}^{K+1} p_k(\boldsymbol{Y}^k | \boldsymbol{X}^k, \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^Y_k, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \boldsymbol{C}^k) p_k(\boldsymbol{X}^k | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X_k, \boldsymbol{C}^k). \tag{1.10}$$

This can be easily shown[3]:

$$\begin{aligned} & P(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \mathbf{C}) \\ & = \int P(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y, \mathbf{C}) \times \\ & \qquad p(\underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y) \, \mathrm{d}(\underset{\sim}{\boldsymbol{\delta}}^X, \underset{\sim}{\boldsymbol{\delta}}^Y, \underset{\sim}{\boldsymbol{\sigma}}^2_X, \underset{\sim}{\boldsymbol{\sigma}}^2_Y) \\ & = \prod_{k=1}^{K+1} \int p_k(\boldsymbol{Y}^k | \boldsymbol{X}^k, \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^Y_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \sigma^2_{k,Y}, \boldsymbol{C}^k) p_k(\boldsymbol{X}^k | \underset{\sim}{\boldsymbol{\alpha}}^X_k, \underset{\sim}{\boldsymbol{\delta}}^X_k, \sigma^2_{k,X}, \boldsymbol{C}^k) \times \\ & \qquad p(\underset{\sim}{\boldsymbol{\delta}}^X_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \sigma^2_{k,X}, \sigma^2_{k,Y} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X_k, \underset{\sim}{\boldsymbol{\alpha}}^Y_k) \; \mathrm{d}(\underset{\sim}{\boldsymbol{\delta}}^X_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \sigma^2_{k,X}, \sigma^2_{k,Y}) \\ & = \prod_{k=1}^{K+1} \int p_k(\boldsymbol{Y}^k | \boldsymbol{X}^k, \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^Y_k, \underset{\sim}{\boldsymbol{\delta}}^Y_k, \underset{\sim}{\boldsymbol{\beta}}, \delta^Y_{10}, \sigma^2_{k,Y}, \boldsymbol{C}^k) p(\underset{\sim}{\boldsymbol{\delta}}^Y_k, \beta_k, \sigma^2_{k,Y} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^Y_k) \, \mathrm{d}(\underset{\sim}{\boldsymbol{\delta}}^Y_k, \sigma^2_{k,Y}) \end{aligned}$$

[3]In the following, $\underset{\sim}{\boldsymbol{\delta}}^X$ includes the exposure model intercepts, but $\underset{\sim}{\boldsymbol{\delta}}^Y$ includes only the coefficients of the covariates.

$$\int p_k(\boldsymbol{X}^k|\underset{\sim}{\boldsymbol{\alpha}}_k^X, \underset{\sim}{\boldsymbol{\delta}}_k^X, \sigma_{k,X}^2) p(\underset{\sim}{\boldsymbol{\delta}}_k^X, \sigma_{k,X}^2|\boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}_k^X)\ \mathrm{d}(\underset{\sim}{\boldsymbol{\delta}}_k^X, \sigma_{k,X}^2)$$

$$= \prod_{k=1}^{K+1} p_k(\boldsymbol{Y}^k|\boldsymbol{X}^k, \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}_k^Y, \delta_{k0}^Y, \beta_k, \boldsymbol{C}^k) p_k(\boldsymbol{X}^k|\underset{\sim}{\boldsymbol{\alpha}}_k^X, \boldsymbol{C}^k) \tag{1.11}$$

Note that all likelihoods in (1.11) are marginal densities of linear regression models over the regression coefficients and variance terms with Normal-Inverse Gamma priors. Raftery et al. (1997) provided closed form calculations of this marginal likelihood. However, this calculation requires the inversion of a matrix with dimension equal to the number of observations, and is computationally intensive. Since the marginal likelihood is only used in the calculation of Bayes factors, we approximate the Bayes factors when necessary using the BIC (Raftery, 1995).

Sampling all model parameters using MCMC

Sampling the regression coefficients and residual variance terms It is worth noting that centering the covariates C_j is an important component of LERCA, since it allows the outcome model intercepts δ_{k0} to depend solely on δ_{10}, β_k and $\boldsymbol{s}$, and not on δ_{kj}^Y. This simplifies the form of the full conditional distribution for many coefficients.

We update coefficients δ_{kj}^X for which $\alpha_{kj}^X = 0$ separately from the ones with $\alpha_{kj}^X = 1$. Parameters δ_{kj}^X for which $\alpha_{kj}^X = 0$ are set to 0. Let $j_1, j_2, \ldots, j_{N_x}$ be the indices such that $\alpha_{kj_l} = 1$, $l = 1, 2, \ldots, N_x$. Then,

$$(\delta_{k0}^X, \delta_{kj_1}^X, \delta_{kj_2}^X, \ldots, \delta_{kj_{N_x}}^X)^T|\text{Data}, \bullet \sim MVN_{N_x+1}(\mu_X, \Sigma_X),$$

$$\text{where } \Sigma_X = \left(\frac{1}{\sigma_{k,X}^2}\tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}} + \frac{1}{\sigma_0^2}I_{N_x+1}\right)^{-1} \text{ and } \mu_X = \Sigma_X\left(\frac{1}{\sigma_{k,X}^2}\tilde{\boldsymbol{V}}^T\boldsymbol{X}^k + \frac{1}{\sigma_0^2}\tilde{\mu}_0\right)$$

where $\tilde{\boldsymbol{V}} = (\boldsymbol{1}, \boldsymbol{C}_{j_1}^k, \boldsymbol{C}_{j_2}^k, \ldots, \boldsymbol{C}_{j_{N_x}}^k)$ is the design matrix of data in experiment k based on the included covariates, and $\tilde{\mu}_0$ is a vector of length $N_x + 1$ of repeated values μ_0.

The full conditional distribution of the variance term $\sigma_{k,X}^2$ is also of known form

$$\sigma_{k,X}^2|\text{Data}, \bullet \sim IG(a_X, b_X),$$

$$\text{where } a_X = a_0 + \frac{n_k}{2},\ \ b_X = b_0 + \frac{1}{2}(\boldsymbol{X}^K - \boldsymbol{V}\underset{\sim}{\boldsymbol{\delta}}_k^X)^T(\boldsymbol{X}^k - \boldsymbol{V}\underset{\sim}{\boldsymbol{\delta}}_k^X),$$

where n_k is the number of observations in experiment k, and $\boldsymbol{V} = (\boldsymbol{1}, \boldsymbol{C}^k)$. The full conditional posterior distributions of $\delta_{kj}^Y, \sigma_{k,Y}^2$ are similar and are omitted.

The parameter δ_{10}^Y is included in the mean structure of the outcome model for all experiments. Its full conditional posterior distribution is $\delta_{10}^Y|\text{Data}, \bullet \sim N(\mu, \sigma^2)$ where

$$\sigma^2 = \left(\frac{1}{\sigma_0^2} + \sum_{k=1}^{K+1} \frac{n_k}{\sigma_{k,Y}^2} \right)^{-1}$$

and

$$\mu = \sigma^2 \left[\frac{\mu_0}{\sigma_0^2} + \sum_{k=1}^{K+1} \frac{1}{\sigma_{k,Y}^2} \sum_{i \in g_k} \left(y_i - \sum_{l=1}^{k-1} \beta_l (s_l - s_{l-1}) - \beta_k (x_i - s_{k-1}) - \sum_{j=1}^{p} \delta_{kj}^Y C_{ij} \right) \right],$$

where $\sum_a^b = 0$ if $b < a$. Similarly, the full conditional posterior distribution of β_k uses data from experiments $k, k+1, \ldots, K+1$, and is $\beta_k|\text{Data}, \bullet \sim N(\mu, \sigma^2)$ where

$$\sigma^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_{k,Y}^2} \sum_{i \in g_k} (x_i - s_{k-1})^2 + (s_k - s_{k-1})^2 \sum_{l=k+1}^{K+1} \frac{n_l}{\sigma_{l,Y}^2} \right)^{-1}$$

and

$$\begin{aligned} \mu = & \sigma^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma_{k,Y}^2} \sum_{i \in g_k} (x_i - s_{k-1}) \big(y_i - \delta_{k0}^Y - \textstyle\sum_j \delta_{kj}^Y C_{ij} \big) + \right. \\ & \left. \sum_{l=1}^{K+1} \frac{1}{\sigma_{l,Y}^2} \sum_{i \in g_l} (s_k - s_{k-1}) (y_i - \delta_{k0}^Y - \textstyle\sum_{e=k+1}^{l-1} \beta_e (s_e - s_{e-1}) - \beta_l (x_i - s_{l-1}) - \sum_j \delta_{lj}^Y C_{ij}) \right) \end{aligned}$$

Sampling the experiment configuration and inclusion indicators The experiment configuration and inclusion indicators can be updated separately, or simultaneously. We first describe the separate update of $\boldsymbol{s}$ and $(\underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y)$, and afterwards we will discuss why occasional simultaneous sampling was deemed necessary. One of the three moves (separate, jump over, jump within) is performed at every iteration with probability $0.8, 0.1$, and 0.1 accordingly.

(separate) The experiment configuration and inclusion indicators are updated separately and conditionally on each other. For the update of the experiment configuration $\boldsymbol{s}$, the full conditional likelihood in (1.9) is used. k is chosen uniformly over $\{1, 2, \ldots, K\}$ and $s^* \sim U(s_{k-1}, s_{k+1})$ is drawn. Alternatively, s^* could be sampled from a truncated normal distribution centered at s_k. If s^* violates $s_{k+1} - s^* \geq d_k$ or $s^* - s_{k-1} \geq d_{k-1}$, the move is automatically rejected.

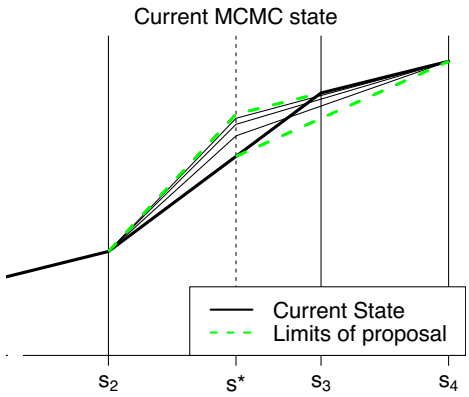


Figure 1.10: Values of β_k, β_{k+1} for the separate move shown in Figure 1.4 are proposed such that the estimated ER are within the limits shown in dashed green lines. The black solid line correspond to the current state of the ER.

Otherwise, the move $\boldsymbol{s} \to \boldsymbol{s}^* = (s_1, s_2, \ldots, s_{k-1}, s^*, s_{k+1}, \ldots, s_K)$ is proposed with all other parameters (excluding $\underset{\sim}{\boldsymbol{\beta}}$) fixed to their current values. New values of $\underset{\sim}{\boldsymbol{\beta}}$ are necessary to ensure that the ER is continuous at the proposed state. All coefficients but β_k, β_{k+1} are fixed to their current values, and new values for β_k, β_{k+1} are proposed such that the intercepts of the adjacent experiments are also fixed. If $s^* < s_k$, the proposed value β^*_{k+1} is sampled from a uniform distribution between the values β_{k+1} (current state) and

$$\tilde{\beta}_{k+1} = (s_{k+1} - s^*)^{-1}\big(\delta^Y_{(k+2)0} - \delta^Y_{k0} - \beta_k(s^* - s_{k-1})\big),$$

where $\tilde{\beta}_{k+1}$ is the slope that would connect the value of the ER at point s_{k+1} with the value of the ER at point s^* at the current state. Figure 1.10 shows the the limits of the proposed ER. Based on the sampled value for β^*_{k+1}, the proposed value for β_k is

$$\beta^*_k = (s^* - s_{k-1})^{-1}\big(\delta^Y_{(k+2)0} - \delta^Y_{k0} - \beta^*_{k+1}(s_{k+1} - s^*)\big).$$

Similarly for $s^* > s_k$ by sampling β^*_k from a uniform that has similar properties.
Since the likelihood factorizes as shown in (1.9) the likelihood ratio of the Metropolis-Hastings acceptance probability includes terms only for experiments $k, k+1$. The prior ratio includes terms for the experiment configuration distribution in (1.8), and the prior for β_k, β_{k+1}. If a uniform distribution is used to sample s^*, the proposal for the cutoffs is symmetric, and the proposal ratio corresponds to the proposal ratio for coefficients β_k, β_{k+1}. This is equal to $|\beta_{k+1} - \tilde{\beta}_{k+1}|/|\beta^*_k - \tilde{\beta}^*_k|$, where β^*_k is the proposed value and $\tilde{\beta}^*_k$ is

the one boundary of the proposal distribution for β_k in the reverse move.

After we accept or reject the move $\boldsymbol{s} \to \boldsymbol{s}^*$, we update the inclusion indicators based on their full conditional (1.6), which we show here. Let A^* be all parameters but α_{kj}^X, α_{kj}^Y and δ_{kj}^Y. For $\alpha \in \{0, 1\}$

$$\begin{aligned}
p(\alpha_{kj}^Y = \alpha | \text{Data}, A^*, \alpha_{kj}^X) =& \frac{p(\delta_{kj}^Y = 0, \alpha_{kj}^Y = \alpha | \text{Data}, A^*, \alpha_{kj}^X)}{p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha, \text{Data}, A^*, \alpha_{kj}^X)} \\
=& \frac{p(\text{Data}, A^* | \delta_{kj}^Y = 0, \alpha_{kj}^Y = \alpha, \alpha_{kj}^X) p(\delta_{kj}^Y = 0, \alpha_{kj}^Y = \alpha | \alpha_{kj}^X)}{p(\text{Data}, A^* | \alpha_{kj}^X) p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha, \text{Data}, A^*, \alpha_{kj}^X)} \\
\propto& \frac{p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha, \alpha_{kj}^X) p(\alpha_{kj}^Y = \alpha | \alpha_{kj}^X)}{p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha, \text{Data}, A^*, \alpha_{kj}^X)} \\
=& \frac{p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha) p(\alpha_{kj}^Y = \alpha | \alpha_{kj}^X)}{p(\delta_{kj}^Y = 0 | \alpha_{kj}^Y = \alpha, \text{Data}, A^*)}, \ \alpha \in \{0, 1\}, \qquad (1.12)
\end{aligned}$$

where the numerator consists of the product of two prior probabilities, and the denominator consists of the posterior probability that $\delta_{kj}^Y = 0$. This has been seen previously in a different context (Antonelli et al., 2017a), and consists a computational improvement over previous implementations of this prior distribution that utilized the MC^3 algorithm (Madigan et al., 1995; Wang et al., 2012).

However, sampling the inclusion indicators and experiment configuration separately can lead to slow convergence. For example, consider our simulation scenario where the true experiment configuration is $(2, 4, 7)$, and starting values randomly set to $(0.5, 2, 7)$. Based on the separate move point s_1 is always proposed to be updated between $s_0, s_2 = 2$, which can lead to slow mixing. The jump over and jump within moves are meant to alleviate such issues. In both situations, sampling of $\boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}}$ is performed using the marginalized likelihood (1.10)

$$p(\boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}} | \text{Data}, \delta_{10}^Y) \ \propto \ p(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}}, \delta_{10}^Y, \boldsymbol{C}) \, p(\boldsymbol{s}) \, p(\underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y) \, p(\underset{\sim}{\boldsymbol{\beta}}).$$

Integrating all other parameters out allows us to perform sampling of the experiment configuration without heavy fine tuning of proposal distributions.

(jump over) This move is designed to alleviate the MCMC issue described above by proposing a simultaneous move of $(\boldsymbol{s}, \underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y, \underset{\sim}{\boldsymbol{\beta}})$. $k \in \{1, 2, \ldots, K\}$ is again chosen uni-

formly, but now a new location of the experiment configuration s^* is generated uniformly over $(s_0, s_{K+1}) \setminus [s_{k-1}, s_{k+1}]$ (necessarily not between s_k, s_{k+1}). The move $\boldsymbol{s} \to \boldsymbol{s}^* = (s_1, s_2, \ldots, s_{k-1}, s_{k+1}, \ldots, s_{j-1}, s^*, s_j, \ldots, s_K)$ proposes a combination of experiments $k, k+1$ and a split in some randomly chosen experiment j. For example, in Figure 1.4(d), the proposed move splits the first experiment (s_0, s_1) in two $(s_0, s^*), (s^*, s_1)$, and combines the experiments $(s_2, s_3), (s_3, s_4)$.

The inclusion indicators of the unchanged experiments remain to their current values, but new values need to be proposed for the combined or split experiments. If in the current state a variable is included in the model of both experiments, the proposed inclusion indicators should reflect the intuition that the variable is potentially important in the model of the combined experiments. Therefore, the inclusion of a variable in the combined experiment is proposed with very low, mediocre and very high probability if none, one or both of the current experiments include it. The values chosen were $(0.01, 0.5, 0.99)$ accordingly. Similarly, two sets of inclusion indicators need to be proposed for the split experiment. A variable is proposed to be included in the model of one of the two experiments with lower and higher probability if the variable was included in the initial model or not. The values chosen were $(0.2, 0.95)$.

Values for $\underset{\sim}{\boldsymbol{\beta}}$ are proposed to ensure that the proposed state corresponds to a continuous ER. Unchanged experiments remain the same. Experiments are combined by connecting the edges of the two linear segments, and values of the split experiments are proposed using a normal perturbation of the current value with variance σ^2_{tune}. Figure 1.4(d) shows proposed states of the ER.

The move is accepted or rejected with probability equal to the product of the following:

1. The likelihood ratio for split and combined experiments approximated using the BIC for the exposure model and the outcome model (regressing $\boldsymbol{Y}^k - (\mathbb{1}, \boldsymbol{X}^k - s_{k-1}\mathbb{1})(\delta^Y_{k0}, \beta_k)^T$ on $\boldsymbol{C}^k$ without an intercept).

2. The prior ratio for the experiment configuration (1.8), and for the inclusion indicators (1.4) and the coefficients β_k for the combined and split experiments.

3. The proposal ratio for $\boldsymbol{s}$, β_k and $(\underset{\sim}{\boldsymbol{\alpha}}^X, \underset{\sim}{\boldsymbol{\alpha}}^Y)$

$$\frac{(s_{K+1} - s_0) - (s_j - s_{j-1})}{(s_{K+1} - s_0) - (s_{k+1} - s_{k-1})} \exp\left\{\frac{u^2 - u^{*2}}{2\sigma^2_{tune}}\right\} \prod_{\substack{l\in\{0,1,2\}\\ m\in\{0,1\}}} (p^c_{lm})^{n^s_{ml} - n^c_{lm}} \prod_{\substack{l\in\{0,1\}\\ m\in\{0,1,2\}}} (p^s_{lm})^{n^c_{ml} - n^s_{lm}},$$

where p^c_{lm} is the probability that of proposing $\alpha = m \in \{0, 1\}$ in the combined experiment when $l \in \{0, 1, 2\}$ of the two initial experiments had $\alpha = 1$, p^s_{lm} is the probability of proposing $\alpha = 1$ in $m \in \{0, 1, 2\}$ of the two experiments when the initial experiment chosen to be split had $\alpha = l \in \{0, 1\}$, and n^c_{lm}, n^s_{lm} is the number of times that each event occurred when moving from the current to the proposed state. Lastly, u is the difference of the slope for the experiment that was split from the slope of the first split experiment in the proposed state, and u^* is the difference of the slope in the first of the experiment that is combined from the slope of the combined experiment in the proposed state.

(jump within) This move is similar to the "jump over" but maintaining the ordering of the locations in $\boldsymbol{s}$. $k \in \{1, 2, \ldots, K\}$ is again chosen uniformly, and a new value s^* is proposed within the interval (s_{k-1}, s_{k+1}). New values for the coefficients β_k, β_{k+1} are proposed as in the separate move. New values of the inclusion indicators are also proposed for the experiments $k, k + 1$. In fact, C_j is proposed to be included in the outcome model of an experiment with high probability if both current models include it, mediocre probability if only one of the models include it, and low probability if none of the models include it. Similarly for the inclusion indicators of the exposure model. The acceptance probability of this move is similar to the one described above, and is omitted here.

MCMC convergence

One quantity that we use for convergence inspection is the mean exposure response curve calculated over a set of exposure values within the exposure range. Such a set might be an equally spaced grid of points over the interval (s_0, s_{K+1}), denoted by $\mathcal{G}$. For each value $x \in \mathcal{G}$ and MCMC iteration t identify the experiment $k = k_t(x)$ that x belongs to. Then, for observation i calculate the expected response at value x, by defining $\tilde{w}_i(x) = (1, x, C_{i1}, \ldots, C_{ip})^T$ and calculating $\widehat{Y}_{it}(x) = \tilde{w}_i(x)^T \gamma_{kt}$ where γ_{kt} is the posterior sample of

$(\delta^Y_{k0}, \beta_k, \delta^Y_{k1}, \ldots, \delta^Y_{kp})^T$ in iteration t. Finally, the t-posterior sample of the mean response at point $x \in \mathcal{G}$ is the average of the expected responses over the individuals in the sample $\widehat{Y}_t(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_{it}(x)$.

Convergence could be examined by visual inspection of trace plots of $\widehat{Y}(x)$ for all $x \in \mathcal{G}$. Based on multiple chains of the MCMC, we calculate the potential scale reduction factor (PSR) for the mean response at every point $x \in \mathcal{G}$ (Gelman and Rubin, 1992). We consider that the MCMC has converged if $|\text{PSR} - 1| < c$ for all $x \in \mathcal{G}$. An alternative quantity based on which MCMC convergence can be examined is $\hat{\Delta}(x) = \beta_{k_x}$.

1.8.4 Simulating data with differential confounding at different exposure levels

In simulation studies, data are most often simulated in the following order: covariates $C_1, C_2, \ldots, C_p$, exposure X given a subset of $C_1, C_2, \ldots, C_p$, and outcome Y given X and a potential different subset of $C_1, C_2, \ldots, C_p$. Data with differential confounding at different exposure levels could imply, in its most generality, that the exposure X is generated with different predicting variables at different exposure levels. Generating data with such structure is complicated since the actual X values define the exposure level that an observation belongs to, and the exposure level in which an observation belongs to defines the set of predictors. For that reason, instead of following the $\boldsymbol{C}, X|\boldsymbol{C}$ approach to data simulation, we generate the exposure values X first, and $\boldsymbol{C}$ is generated conditional on X, ensuring that the target experiment-specific mean and variance of $X, \boldsymbol{C}$, and correlation of all variables remain the same, as if the data were generated with the typical $\boldsymbol{C}, X|\boldsymbol{C}$ order. Generating the outcome with different predictors at different exposure model is straightforward by including terms of the form $\delta^*_j C_j I(X \geq s_k)$, or by using a separate outcome model within each experiment. In all situations, one should ensure that data are generated in such a way that the true ER is continuous.

The "target" data generating mechanism

Given $K, \boldsymbol{s}$, we would like the exposure X to be generated such as $E(X)$ and $Var(X)$ are controllable quantities, since they are closely related to the exposure range of each experi-

ment, and we would like to to ensure that simulation results are not driven by the inherit variability in X. Furthermore, we would like to ensure that $Var(C_j)$ is approximately the same across experiments and across covariates, such that the the magnitude of δ_{kj}^Y has similar interpretation in terms of correlation.

As discussed above, data $(X, \boldsymbol{C})$ are usually generated in the order $\boldsymbol{C}$ followed by $X|\boldsymbol{C}$, using a model for which $E(X|\boldsymbol{C}) = \delta_0 + \sum_{j=1}^p \delta_j C_p$. Instead of setting target values for δ_j, we set target correlations $Cor(X, C_j)$ and calculate the δ_j's that correspond to these correlations. Alternatively, one could specify target δ_j's ensuring that $Var(X) \geq \sum_{j=1}^p \delta_j^2 Var(C_j)$.

We require that $E(C_j|X = x)$ is continuous in x to ensure that the joint distribution (X, C_j) is realistic, and does not have "jumps" at the points of the experiment configuration.

Based on the above, the following represent the target quantities of our data generation:

- $Var(X), E(X)$ are fixed,
- C_j are independent random variables with known variance within each experiment,
- $Cor(C_j, X)$ are fixed and δ_j can be calculated, using $Cor(X, C_j) = \delta_j \sqrt{\frac{Var(C_j)}{Var(X)}}$,
- The function $E(C_j|X = x)$ is continuous in x.

Ensuring that $E(C_j|X = x)$ is continuous in x across experiments is performed in the following way: Given $Var(X)$, a model for $\boldsymbol{C}|X$ that gives rise to data with the target $Var(C_j), Cor(X, C_j)$ is considered. The variance-covariance targets do not impose any restrictions on the model intercept. For the first experiment, the intercept can be chosen arbitrarily, and for the subsequent experiments intercepts are chosen to ensure that $\lim_{t \to x^-} E(C_j|X = t) = \lim_{t \to x^+} E(C_j|X = t)$ at all points x.

Generating the data set maintaining target quantities

As discussed above, $Cor(X, C_j), Var(C_j)$, and $Var(X)$ are considered known, from which we can derive $Cov(X, C_j)$. We generate data with the following order:

1. X is generated from a distribution with mean $E(X)$, and variance $Var(X)$. In our simulations X is uniform over the exposure range.

2. Taking advantage of the laws of the multivariate normal distribution we generate

$$\begin{aligned}\boldsymbol{C}|X \sim & MVN_p(\bar{\mu}, \bar{\Sigma}), \text{ where} \\ \bar{\mu} = & E(\boldsymbol{C}) + \frac{Cov(\boldsymbol{C}, X)}{Var(X)}(X - EX), \text{ and} \\ \bar{\Sigma} = & \text{diag}(V(\boldsymbol{C})) - \frac{1}{Var(X)} Cov(\boldsymbol{C}, X) Cov(\boldsymbol{C}, X)^T,\end{aligned}$$

where $Cov(\boldsymbol{C}, X) = (Cov(C_1, X), Cov(C_2, X), \ldots, Cov(C_p, X))^T$, and diag$(V(\boldsymbol{C}))$ is a diagonal $p \times p$ matrix with entries $Var(C_j), j = 1, 2, \ldots, p$.

3. The marginal means of each variable C_j within each experiment is calculated by ensuring that the function $E(C_j|X = x)$ which corresponds to the j^{th} entry of the vector $\bar{\mu}$ is continuous at the points of experiment change.

4. Covariates C_j are subtracted their overall mean.

A simple linear regression form is used to generate the outcome within each experiment. In experiment k, the outcome is simulated from a model $Y|X, \boldsymbol{C} \sim N(\xi_{k0} + \xi_{k1}\phi(X) + \sum_{j=1}^{p} \xi_{k(j+1)} C_j, \sigma^2_{k,Y})$, where $\phi()$ is a continuous function, and the residual variance $\sigma^2_{k,Y}$ is set equal across k. We ensure that the true ER function $E(Y|X)$ is continuous in X by appropriately setting the intercept values ξ_{k0}. The intercept in experiment 1 is decided, and for each experiment onwards we set ξ_{k0} such that

$$\lim_{x \to s_k^-} E[Y|X = x] = \lim_{x \to s_k^+} E[Y|X = x] \iff \xi_{(k+1)0} = \xi_{k0} + (\xi_{k1} - \xi_{(k+1)1}) s_k.$$

1.8.5 Additional simulation results

Simulations in the presence of *local* confounding

Table 1.3 shows the correlation of the covariates with the exposure and the coefficients of the covariates in the outcome model for the data simulating scenario with local confounding: different confounders at different levels of the exposure.

Figure 1.11 shows the the root MSE (rMSE) as a function of the exposure value $x \in (0, 10)$. LERCA has the lowest rMSE at the low exposure levels followed by GAM. All methods are comparable for the middle exposure values, and GAM performs slightly better than LERCA at high levels.

Table 1.3: Correlation between the covariates and exposure, and outcome coefficients in each experiment, for scenarios with local confounding.

	Covariate - Exposure				Covariate - Outcome			
	$x \in g_1$	$x \in g_2$	$x \in g_3$	$x \in g_4$	$x \in g_1$	$x \in g_2$	$x \in g_3$	$x \in g_4$
C_1	0.423	0.525	0.402	0	0.641	0	0	0
C_2	0.524	0.572	0	0.503	0.962	0.919	0.593	0.651
C_3	0.522	0	0.447	0	0.646	0.643	0.616	0.58
C_4	0	0.528	0	0	0	0.633	0	0
C_5	0	0	0.533	0.539	0	0	0.658	0
C_6	0	0	0	0.509	0	0	0	0.52
C_7	0	0	0	0	0	0	0	0
C_8	0	0	0	0	0	0	0	0

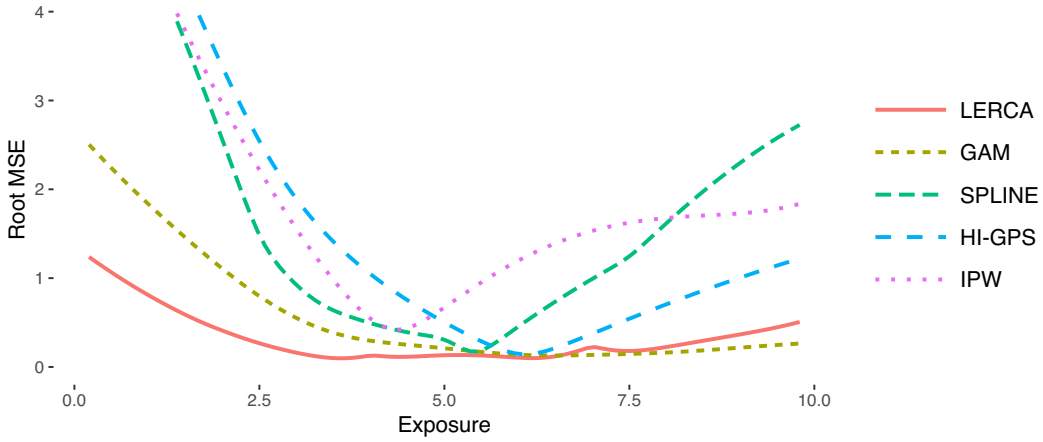


Figure 1.11: Mean Root MSE as a function of the exposure $x \in (0, 10)$.

Simulations in the presence of *global* confounding

Briefly, data are generated with covariates C_1, C_2, C_3 as predictors of exposure and C_2, C_3, C_4 as predictors of the outcome and the adjusted R-squared of the true exposure and outcome models was 0.73 and 0.94 accordingly. Table 1.4 shows the correlation of covariates with the exposure and the outcome model coefficients in the data simu-

Table 1.4: Correlation between the covariates and exposure, and outcome coefficients in each experiment, for scenario with global confounding.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Exposure	0.423	0.524	0.522	0	0	0	0	0
Outcome	0	0.812	0.93	0.82	0	0	0	0

lating scenario with global confounding (same confounders with constant confounding strength across exposure levels) and true quadratic ER. Figure 1.12 shows the estimated ER for each data set and the average estimated ER based on LERCA and alternative methods. In Figure 1.13, the root MSE for all methods is plotted as a function of the exposure $x \in (0, 10)$.

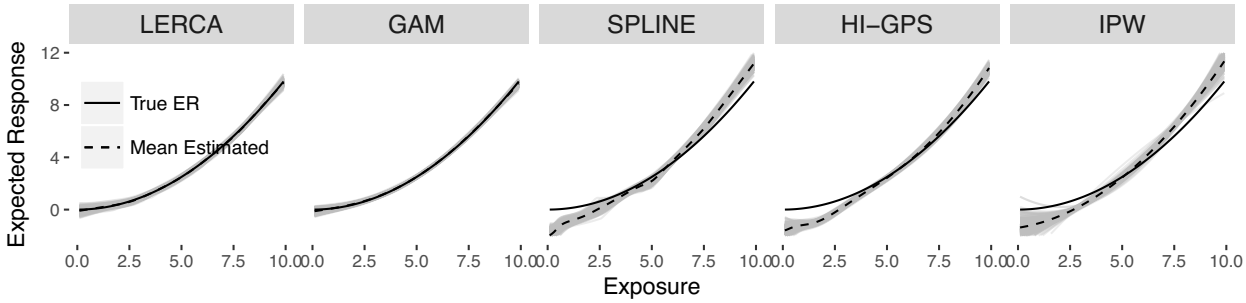


Figure 1.12: Simulation results in the presence of global confounding. Grey lines correspond to estimated ER for each simulated data set, dashed lines correspond to the mean ER over all simulated data sets, and the solid line corresponds to the true ER.

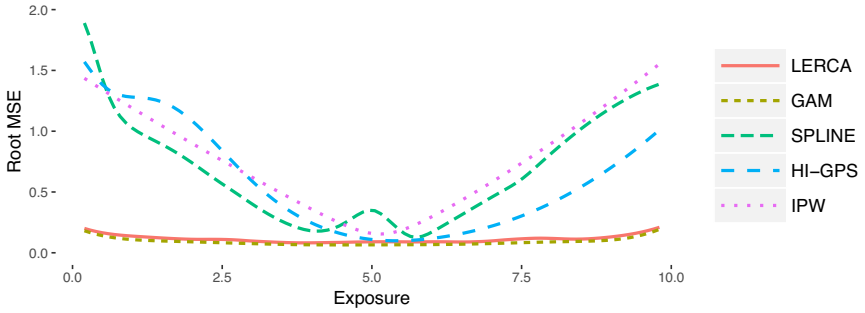


Figure 1.13: Root MSE of all methods in the presence of global confounding as a function of the exposure $x \in (0, 10)$.

Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching

Georgia Papadogeorgou

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Christine Choirat

Department of Biostatistics

Harvard Chan School of Public Health

Corwin M. Zigler

Department of Biostatistics

Harvard Chan School of Public Health

2.1 Introduction

Methods based on propensity score matching are widely used to estimate causal effects with observational data. Such methods rely crucially on the assumption of no unmeasured confounding. In settings of spatially-indexed data, unobserved confounders may exhibit a spatial pattern, inviting the use of spatial information to serve as proxy for similarity of units with respect to unmeasured confounding factors. Methods for confounding adjustment with spatially-indexed data have been most often considered in the context of regression adjustment, as in Paciorek (2010) and in related work that does not target confounding adjustment per se, but has been used for modeling spatially correlated residuals via spatial random effects (Hodges and Reich, 2010; Lee and Neocleous, 2010; Lee and Sarran, 2015; Chang et al., 2013; Congdon, 2013).

In this paper, we unite the use of spatially-indexed data with propensity score matching while preserving the most salient benefits of using propensity scores. These benefits include the explicit comparison of treatments or policy interventions to estimate policy-relevant estimands such as the Average Treatment Effect on the Treated, as well as the oft-cited virtues of propensity score analysis related to the hypothetical "design" of a randomized study, for example, the ability to check observed covariate balance and overlap (Rubin, 2008). Augmenting such benefits with the notion that geographically closer units may exhibit similar unmeasured confounding profiles presents a methodological challenge.

The methods here are motivated by the threat of unmeasured spatial confounding that arises in studies of air pollution, where complex climatological and atmospheric processes are known to vary spatially and have strong associations with ambient air pollution, but are often unmeasured. For example, consider ambient ozone pollution, which has been previously linked to adverse health outcomes (Bell et al., 2004; Jerrett et al., 2009). A variety of regulatory strategies in the U.S. are designed to reduce ambient ozone pollution through incentivizing power-generating facilities (i.e., "power plants") to reduce emissions of nitric oxide and nitrogen dioxides (NO_x). When combined with sunlight and in the presence of available volatile organic compounds, NO_x emissions initiate atmo-

spheric chemical reactions to form ambient ozone pollution (Allen, 2002). What's more, regions where conditions tend to encourage the formation of ozone might be more likely to impose stricter rules on NO_x emissions. Thus, evaluating the effectiveness of emission-control strategies installed at power plants is met with the challenge that complete data on all relevant climatological, atmospheric, and regulatory confounders is almost never available but are expected to vary spatially. The goal of this paper is to employ a matching procedure anchored to the propensity score to investigate whether, among coal or natural gas power plants, installation of selective catalytic or selective non-catalytic (SCR/SNCR) NO_x emission control technologies is more effective than alternatives for reducing ambient ozone. The treatment assignment and outcome of interest are depicted in Figure 2.1. Propensity scores are particularly useful for this type of policy evaluation because of the ability to adjust for confounding without strong reliance on a parametric model and the ability to empirically assess covariate balance and overlap. However, unmeasured spatial confounding presents a strong threat to the validity of a standard propensity-score analysis.

To confront these challenges, we present a new methodology, termed Distance Adjusted Propensity Score Matching (DAPSm) which incorporates information from spatially-indexed data with the known virtues of propensity score matching. DAPSm incorporates observations' spatial proximity into a matching procedure designed to adjust for

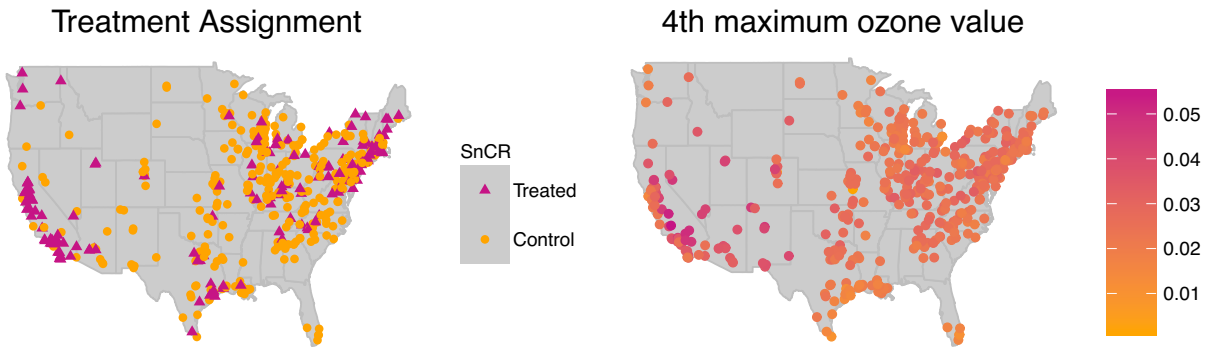


Figure 2.1: Map of facilities, colored by whether they are treated (yellow) or control (red), and map of ozone concentration surrounding power plants.

observed confounders while adjusting for unmeasured spatial confounders by emphasizing, to varying degrees governed by a tuning parameter, the spatial proximity of matches. The central challenge of incorporating spatial proximity into propensity score matching is that proximity is a relative measure between two units, not a unit-specific measure like a confounder or the propensity score itself.

DAPSm shares important commonalities with the recently-proposed work of Keele et al. (2015) in that both are matching methods that aim to leverage spatial proximity of units. We evaluate DAPSm relative to the method of Keele et al. (2015) throughout, but note here that, despite similar conceptual goals, these methods are not directly comparable. The most salient difference has to do with reliance on the propensity score; DAPSm combines spatial proximity with propensity scores, whereas Keele et al. (2015) provides an integer programing method that matches directly on covariates (i.e., not using propensity scores). Our goal here is not to investigate the relative merits of exact vs. propensity score matching, but rather to isolate features related specifically to methods' account of spatial confounding. DAPSm offers a tuning parameter governing the relative prioritization of observed covariate distances (measured through similarity of propensity score estimates) and spatial proximity. Keele et al. (2015) entails a tuning parameter that governs the trade off between spatial proximity of matches and the number of matches selected within a certain tolerance of observed covariate balance.

Both DAPSm and the method of Keele et al. (2015) are evaluated in a simulation study alongside several other reasonable alternatives for incorporating spatial information into propensity score analysis. The methods are then deployed to compare the effectiveness of SCR/SNCR, relative to other strategies, for reducing NO_x emissions and ambient ozone measured across 473 power plants and 921 air pollution monitoring locations in the United States. Ultimately, we show that incorporating spatial information in the matching can lead to substantively different conclusions when evaluating interventions on spatially-indexed observational units.

2.2 Notation, estimand of interest, and outline of propensity score matching

Let Z_i denote the indicator of whether the i^{th} of n observations is subject to treatment, for example, the indicator of whether a power plant is treated with SCR/SNCR ($Z_i = 1$) or not ($Z_i = 0$). Let Y_i be a continuously-scaled outcome, for example, ambient ozone concentration in the area surrounding power plant i.

Each unit or observation is assumed to have two potential outcomes (Rubin, 1974), one under each value of the treatment. We denote $Y_i(0), Y_i(1)$ as the potential outcome of ambient ozone concentration in the area around unit i under value of $z = 0, 1$ respectively. Assuming that the indexing of the observations is done at random, the index i is suppressed. Interest often lies in the estimation of the average treatment effect in the treated population, defined as $ATT = E[Y(1) - Y(0)|Z = 1]$.

Among the assumptions required to estimate the ATT with observational data is that of "ignorable treatment assignment", stating that observed covariates are sufficient to adjust for confounding of the treatment-outcome relationship. More formally, let **C** be a minimal set of confounding variables such as power plant characteristics, weather and atmospheric variables, and area-level demographics. The assumption of ignorability can be stated as:

$$Y(z) \amalg Z|\mathbf{C}, \tag{2.1}$$

under which the ATT can be estimated with observed-data comparisons between outcomes on treated and untreated units, conditional on **C**. Since **C** is assumed minimal, the ignorability assumption in (2.1) does not hold for any strict subset of **C**, implying that observed-data comparisons will not estimate the ATT when conditioning on a strict subset of **C**.

As the dimensionality of **C** increases, investigators often use the propensity score to condense the information in **C** into a "balancing score" that can be used to adjust for confounding when comparing treated and untreated units. The propensity score is defined as the conditional probability of receiving treatment given the covariates, $P(Z_i = 1|\mathbf{C}_i)$. The balancing property of the propensity score (Rosenbaum and Rubin, 1983) implies that

the ignorability assumption in (2.1) can be translated to $Y(z) \amalg Z|P(Z = 1|\mathbf{C})$.

An overview of the various ways in which propensity scores can be used for confounding adjustment can be found in Stuart (2010), but we discuss methods in the context of 1:1 matching without replacement using a caliper. Such a procedure uses propensity score estimates to match one treated unit to one control unit with a similar propensity score estimate. A threshold, called a "caliper," can be used to avoid matching observations with insufficiently similar propensity scores. For example, specifying a caliper of 0.1 prevents the matching of any two observations with propensity scores that differ by more than 0.1 standard deviations of the propensity score distribution. Matching produces a data set of matched treated and control observations with similar propensity score distributions, and thus more similar distributions of the covariates in $\mathbf{C}$ and a treatment indicator that is, under ignorability, unconfounded. The resulting matched data set can be used to estimate the ATT provided that all elements of $\mathbf{C}$ are observed and used to construct the propensity score.

However, it is often the case with observational data that the vector of true confounders $\mathbf{C}$ can be partitioned into two categories, $\mathbf{C} = (\mathbf{X}, \mathbf{U})$, where $\mathbf{X}$ denotes the confounders available in the observed data, and $\mathbf{U}$ denotes confounders that are unobserved. In the presence of unobserved confounders $\mathbf{U}$, the ignorability assumption in (2.1) cannot be satisfied by conditioning solely on the observed $\mathbf{X}$, and the treatment effect is not identifiable from the data.

In many settings it is expected that some elements of $\mathbf{U}$ vary spatially so that locations that are geographically close are similar with regard to $\mathbf{U}$. In this sense, the notion of prioritizing spatial proximity of matches has points of contact with the notion of a spatial "bandwidth" in a geographic regression discontinuity design (such as that in Keele et al. (2015)), where only observations within the bandwidth are regarded as comparable on observed (and unobserved) factors. The method outlined below regards U as unmeasured variables with a distribution over the whole geography of interest that is continuous as a function of space, with closer observations having more similar $\mathbf{U}$.

2.3 Distance Adjusted Propensity Score Matching

We propose a procedure that is anchored to the propensity score for matching on observed confounders, but augments confounding adjustment by incorporating spatial (geographical) information as a proxy for unobserved spatial variables, **U**, such as weather and atmospheric conditions. In the presence of such **U**, prioritizing matched units that are geographically close to each other could yield better covariate balance on all $\mathbf{C} = (\mathbf{X}, \mathbf{U})$, thus (approximately) recovering the ignorability assumption and reducing bias of causal estimates. Formally, the variables $U \in \mathbf{U}$ are such that $\forall \epsilon > 0$ and point s_0 in the geography of interest $\exists \mathcal{N}_\epsilon(s_0)$ open set including s_0 such that $|U(s) - U(s_0)| < \epsilon, \forall s \in \mathcal{N}_\epsilon(s_0)$.
We define the Distance Adjusted Propensity Score (DAPS) as a new quantity for identifying good matches between treated and control units. In contrast to the propensity score which has a value for each unit, the DAPS is defined for every (i, j) pair of treated, control observations. Specifically, for treated unit i and control unit j, the DAPS combines propensity score estimates and relative distances to define: $DAPS_{ij} = w * |PS_i - PS_j| + (1 - w) * Dist_{ij}$, where $w \in [0, 1]$, PS_i, PS_j are propensity score estimates from modeling the treatment conditional on the observed confounders, and $Dist_{ij}$ is a distance measure capturing the proximity of units i, j. DAPS is a weighted average of the propensity score difference used in "standard" propensity score matching and a measure of the distance between treated-control pairs. Therefore, it is a transparent measure of similarity between treated and control units, with an (i, j) pair having small DAPS_{ij} regarded as comparable on the basis of a combination of propensity score difference and spatial proximity.

2.3.1 Choosing the weight w

Setting $w = 1$ corresponds to setting DAPS equal to the absolute propensity score difference, and similarity of treated and control units is based solely on the observed confounders, without regard to spatial proximity. Setting $w = 0$ ignores **X**, defining similarity of units based solely on distance. In practice, w could be specified in the range $[0, 1]$ depending on contextual prioritizaiton of observed confounding and the threats due to any

suspected unobserved spatial confounding, with values closer to 0 for settings where unobserved spatial confounding is of particular concern. Data-driven procedures, such as the one described in Section 2.3.5 can be useful in choosing a w.

2.3.2 Choosing the distance measure

The quantity $Dist_{ij}$ could be specified in many ways to quantify spatial proximity of units i and j. A natural distance measure is the geographical distance between units i, j. A key consideration in choosing a distance measure is that its scale must be made comparable to that of the propensity score to ensure that one quantity does not arbitrarily dominate the calculation of DAPS. Since the absolute propensity score difference of two units can vary across the range $[0, 1]$, the distance measure should also vary between 0 and 1, or on a range similar to the range of estimated propensity score differences. (Alternatively, instead of standardizing $Dist_{ij}$, one could scale w.)

One distance measure we consider is the standardized Euclidean distance (for simulations) or the standardized geo-distance (for the application). Specifically, if $i \in S_t = \{1, 2, \ldots, N_t\}$ is a treated unit, and $j \in S_c = \{1, 2, \ldots, N_c\}$ is a control unit, the standardized distance of i, j is defined as:

$$Dist_{ij} = \frac{d_{ij} - \min_{TC} d}{\max_{TC} d - \min_{TC} d}, \tag{2.2}$$

where d_{ij} is the Euclidean (or geo) distance between i, j, and $\min_{TC} d, \max_{TC} d$ are the minimum and maximum distances of all the treated-control pairs.

Other choices of distance measure can also arise in practice. For example, only permitting matches within certain boundaries (e.g., within states) corresponds to setting $Dist_{ij} = \infty$ for i, j located in different states. Section 2.7.1 presents an alternative definition relying on the empirical CDF of treated-control pairwise distances.

2.3.3 Selecting matches

We provide an R package that performs matching based on DAPS using an optimal or a greedy algorithm. The optimal algorithm uses the `optmatch` R package, and the greedy algorithm is described in Section 2.7.2. Gu and Rosenbaum (1993) found that optimal

matching performed better than greedy matching in returning matched pairs with small Mahalanobis covariate distance, but returned similarly balanced matched data sets.

2.3.4 Specifying Calipers

In DAPSm, a caliper can be defined as the number of DAPS standard deviations beyond which a value of DAPS is deemed too large to produce an appropriate match. In this situation, a treated-control pair cannot be matched if the corresponding DAPS of the pair is larger than the caliper. That is, the caliper is directly applied to the entire DAPS quantity. Calipers could be alternatively defined to pertain separately to each component of DAPS. For example, one type of caliper could prevent any match with propensity score difference exceeding some threshold regardless of DAPS value, with an analogous caliper defined only for distance.

Note that when a caliper is not used, there is an equivalence between DAPSm with $w = 1$ and standard 1-1 nearest neighbor propensity score matching. When a caliper is specified these procedures may not be exactly equivalent due to the definition of the caliper for the two procedures. Standard matching uses the standard deviation of propensity score estimates, while DAPSm uses the standard deviation of DAPS $\overset{w=1}{=}$ |PS difference|.

2.3.5 Data-driven choice of w

In DAPS, there is a transparent interplay between distance of observed covariates (as measured through the difference in the propensity score estimates) and distance of matched pairs. Automated data-driven procedures may be useful for selecting an appropriate value of w. We implement an automated procedure that re-calculates DAPS and performs matching across a range of possible w. As w increases, balance of the observed covariates can be assessed, and the smallest value of w that maintains the absolute standardized difference of means (ASDM) of the observed confounders below a pre-specified cutoff is used. A different balance criterion can also be used. This choice of w assigns the largest possible weight to proximity (and, by extension, to the unmeasured spatial confounders), while still maintaining balance of the observed confounders.

Even though this choice of w is such that it ensures observed covariate balance with re-

spect to a specific criterion, w can be specified alternatively if subject-matter knowledge is available on an unmeasured spatial confounder. For example, there may be a known but unmeasured confounder (e.g., volatile organic compounds, baseline NO_x emissions in an air pollution study) that is regarded as more important than any measured variable. The value of w could be chosen such that DAPSm prioritizes spatial proximity of matched pairs to maximize the chance of balancing the unmeasured spatial confounder, even at the cost of balance on observed covariates. Ability to make such a judgment transparently is a key feature of DAPSm.

2.4 Simulation and Comparison with Alternatives

We conduct a simulation study to explore the performance of DAPSm and several reasonable alternatives for incorporating spatial information, with a focus on how different methods perform across a variety of unmeasured spatial confounding settings, as dictated by the spatial surface of simulated unmeasured confounding. We evaluate methods with respect to mean squared error (MSE) of ATT estimates, balance of observed and unobserved confounders, and number of matches. Data are simulated across the locations of 800 power generating facilities to represent a realistic spatial patterning of units reflecting that of the study of power plant emissions and ozone. Specifically, for each simulated data set, each of 800 fixed locations are simulated to have one unmeasured confounder U generated as a Gaussian Process with Matérn correlation function, four observed confounders $X_i, i = 1, 2, 3, 4$ uncorrelated with U, binary treatment Z, and continuous outcome Y. The specifics of the data generating mechanism can be found in Section 2.7.3.
The Matérn correlation function of the spatial confounder is governed by two parameters, the smoothness, ν, and range, r. The range, r, measures how quickly the correlation of U between two locations decays with distance. When ν is small, the spatial process is rough, and when it is large the process is smooth. See Minasny and McBratney (2005) for a detailed description. Section 2.7.4 shows four generated surfaces of a spatial variable with Matérn correlation function for combinations of small and large values of smoothness and range.

The situation presented here assumes that U is uncorrelatd with $\mathbf{X}$ to highlight the impact of completely unobserved spatial confounders. Situations where U is simulated to have correlation with $\mathbf{X}$ produce similar results with less pronounced gains of incorporating spatial information.

2.4.1 Methods for Comparison with DAPSm

We consider alternative approaches belonging to two general strategies for incorporating spatial information with propensity scores: a) incorporating spatial information in the matching procedure, and b) incorporating spatial information in the propensity score estimates themselves. The former methods estimate the propensity score using $\mathbf{X}$, then perform matching based in part on distance, as done in DAPSm. The latter methods estimate propensity scores that vary according to a spatial pattern by construction, then matches on these "spatial propensity scores". After matching is performed, ATT estimates are acquired through a difference in means of the matched pairs. Further regressions adjustments could be performed in practice.

The previously described method of Keele et al. (2015) is one method that incorporates spatial information in the matching. Even though Keele et al. (2015) advocate for matching directly on covariates, in the simulation study this method was implemented performing exact matching on 5 categories of the propensity score, such that any difference in performance could be attributed solely to the methods' ability to adjust for unmeasured spatial confounding. Simulation results for this method implemented to match directly on covariates are shown in Section 2.7.5.

We further considered another method that incorporates spatial information into the matching procedure, which we refer to as "Matching within Distance Caliper". For this method, a distance caliper is chosen as the maximum distance of potential matched pairs. Within the distance caliper, matching is performed based solely on the propensity score estimated with $\mathbf{X}$. A caliper on the propensity score can be used in addition.

Methods that incorporate spatial information into the propensity score estimates include parametric and non-parametric incorporation of spatial information in the treatment assignment model. A simple approach is the introduction of fixed effects for locations'

latitude and longitude coordinates in the propensity score, in addition to the observed covariates. We refer to this as "Naïve with Coordinates". A more flexible extension is to use Gradient Boosting Models (GBM; Friedman (2001)) to estimate the propensity score, including the coordinates and the observed covariates $\mathbf{X}$. Estimation of the model is performed using the *gbm* R package (Ridgeway, 2007).

While the "Naïve with Coordinates" and GBM approaches are not spatial methods per se, an alternative approach is to augment the propensity score model with a spatial random effect, as implemented using the *spBayes* R package (Finley et al., 2007). Specifically, the propensity score is estimated by fitting $P(Z = 1|\mathbf{X}) = f(\mathbf{X}, W; \theta_W), \quad W \sim GP(0, C)$, where $C = C(\lambda)$ is the spatial random effect correlation matrix with parameters λ. Such approach was not pursued in detail here due to its computational intensity and its poor performance in initial investigations.

We compare the propensity score matching with spatial information methods to the gold standard which uses the data generating outcome model, and the gold standard propensity score ("Gold PS") which uses the true propensity score model conditional on $\mathbf{X}$ and U. Finally, the naïve approach performs propensity score matching using estimates from a model solely on the observed confounders $\mathbf{X}$.

All methods are implemented with 1-1 nearest neighbor optimal matching without replacement. For DAPSm, we present results for the definition of standardized distance defined in (2.2). For GBM, we considered 3^{rd} degree interactions. Results for Matching within Distance Caliper are presented with the distance caliper equal to the 10^{th} percentile of pairwise treated-control distances (the method indicated sensitivity to the choice of distance caliper, other specifications were considered, but are not shown here). The method of Keele et al. (2015) was implemented across a range of values for λ, representing different compromises between the number of returned matches and the distance between matched pairs. As Keele et al. (2015) do not provide specific guidance on the selection of λ, we present results for two values meant to represent two different points in the space of compromises between distance and the number of matches: $\lambda = 0.382$, the median pairwise distance of treated and control units (as done in Keele et al. (2015)), and $\lambda = 0.05$ which was determined in simulation to yield fewer matches and lower MSE for

this range of simulation scenarios. For implementing DAPSm, w was chosen based on the algorithmic procedure described in section 2.3.5.

2.4.2 Simulation Results

Figure 2.2 shows the relative MSE of the effect estimates calculated with a subset of the methods with respect to the Gold PS and for different specifications of smoothness and range of U. MSE is calculated over the subset of simulated data sets for which each method returned matches. Table 2.1 describes the percentage of simulated data sets for which no matching was achieved for each of the methods. As expected, the naïve approach has the highest relative MSE ranging from 24.4 to 46.6. Relative MSE for the gold standard varied from 0.16 to 0.32, indicating that specifying the correct outcome model is more efficient than using the correctly specified propensity score. These approaches did not indicate patterning when varying the spatial structure. Relative MSE for the Naïve

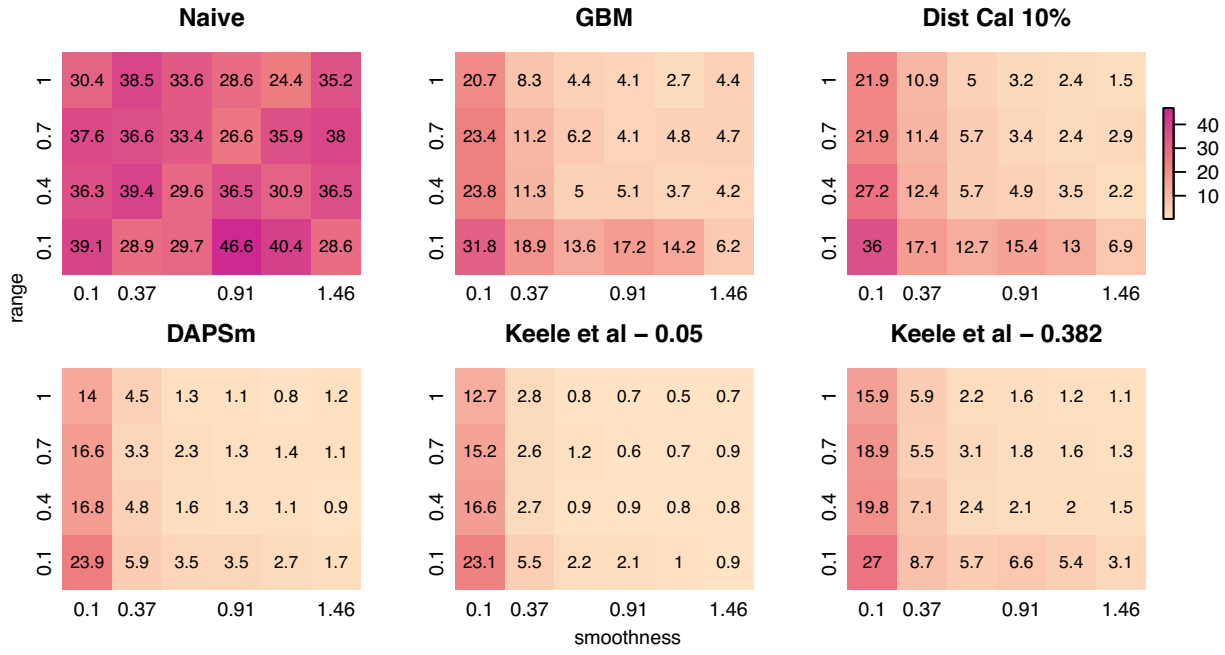


Figure 2.2: Estimates of relative mean squared error over 100 simulated data sets for each specification of smoothness ν (x-axis) and range r (y-axis) for the Matérn correlation function of the unobserved confounder. The baseline MSE corresponds to the Gold PS. Printed values are rounded to the first decimal.

Table 2.1: Percentage of simulated data sets that each method returned no matches (% fail), average number of treated units that were dropped when matches where returned (Dropped), the Interquartile range of number of dropped treated units (IQR), and average distance of matched pairs (Distance).

	Gold PS	Naïve	N.Coords	GBM	DistCal 10%	DAPSm	Keele-0.05	Keele-0.382
% fail	1.5	0.5	0.96	27.67	33.29	0.04	0	0
Dropped	0.06	0.06	0.06	0.23	0	0	55.98	2.11
IQR	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(50,62)	(0,3)
Distance ($\times 100$)	37.4	40.5	36.4	36.1	8.4	2.6	1.9	3.7

with coordinates ranged from 6 to 41.3, and indicated similar patterning as the other spatial methods with respect to the smoothness and range of U, but performed worse in terms of MSE than its more flexible form (GBM) and is emitted from Figure 2.2.

For all methods incorporating spatial information, relative MSE decreases as the surface gets smoother (larger values of smoothness ν) or the spatial correlation remains positive at longer distances (larger values of range r). Similar results are observed for the absolute bias. Among the methods considered based on propensity scores, DAPSm had the lowest MSE across all specifications of range and smoothness, apart from the method of Keele et al. (2015) when $\lambda = 0.05$ was chosen such that it reduces simulation-based MSE compared to $\lambda = 0.382$. Results for the method of Keele et al. (2015) implemented to match directly on the observed covariates, instead of the propensity score, can be found in Section 2.7.5, and showed lower relative MSE than the methods presented here.

We also evaluated methods with respect to the balance of observed and unobserved covariates. Figure 2.3 shows the standardized difference of means of X_1, X_2, U (balance of X_3, X_4 was similar to the balance of X_1, X_2) for the scenarios where $\nu = r = 0.1$ (rough uneven surface), and $\nu = 1.46, r = 1$ (smooth surface). First, the full data ASDM shows that all variables were imbalanced in most simulated data sets. Using the correctly specified propensity score model (Gold-PS) achieved balance of all confounders at the 0.1 cutoff. The naïve approach does not incorporate any spatial information, and the unobserved confounder remains imbalanced. Incorporating coordinates in the estimation of the propensity score improves on balancing U, especially in smoother surfaces. Matching within Distance Caliper performed similarly to Naïve with coordinates in the rough surface. In smooth surfaces Matching within Distance Caliper performed well in balancing

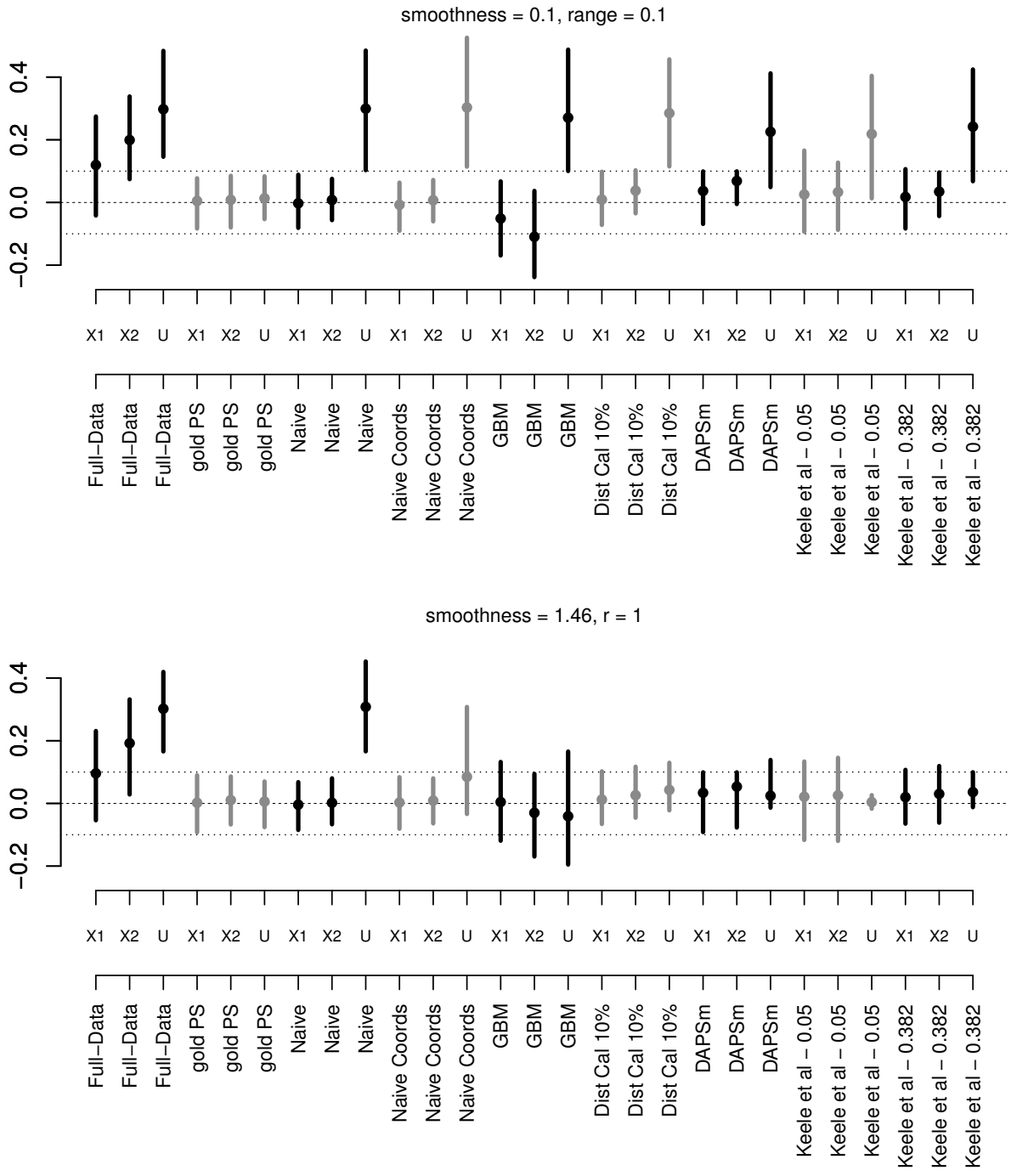


Figure 2.3: Average and (2.5%, 97.5%) intervals of standardized difference of means of (X_1, X_2, U) for the scenarios where $\nu = r = 0.1$ (top), and $\nu = 1.46, r = 1$ (bottom) over 100 Monte Carlo simulations.

both $\mathbf{X}$ and U, although the balance of U is sensitive to the choice of the distance cutoff. The GBM approach exhibited poor balance for all covariates in both scenarios. DAPSm with weight w chosen as described in section 2.3.5 balanced all observed covariates, while improving balance of U in both rough and smooth surfaces. The method of Keele et al. (2015) for $\lambda = 0.382$ returned matches for which balance of the observed variables was better than for $\lambda = 0.05$, but $\lambda = 0.05$ returned matched data sets with better balance on U.

Lastly, the methods considered exhibited substantial variability in terms of the number of achieved matches. Optimal matching algorithms often return no matched pairs. Table 2.1 shows the percentage of simulated data sets that each method failed to return any matches, and the average and IQR of number of treated units that were dropped when matching was achieved. GBM and matching in distance calipers had a high probability of failing to return matches, but when matching was achieved, they failed to match, on average, less than 1, or 0 treated units accordingly. DAPSm failed to return matches for 0.04% of simulated data sets, but matched all treated units otherwise. On the other hand, Keele et al. (2015) returned matches with a significant amount ($\lambda = 0.05$) or a small number ($\lambda = 0.382$) of dropped treated units. Differences in the number of obtained matches should be viewed in light of the fact that confining effect estimation to subsets of the available data can change the causal estimand of interest.

2.5 Comparing the effectiveness of SCR/SNCR emission reduction technologies for reducing NO_x emissions and ambient ozone

Regulatory strategies impacting U.S. power plants are predicated on the knowledge that reducing NO_x emissions reduces ambient ozone, prompting many policies that incentivize the installation of emission control technologies at power plant smokestacks. While many technologies are available, Selective Catalytic Reduction (SCR), and Selective Non-Catalytic Reduction (SNCR) technologies are believed to be among the most efficient for reducing NO_x. However, no study has, to our knowledge, empirically compared

the effectiveness of these strategies to evaluate whether the supposed efficiency gains of SCR/SNCR for reducing NO_x emissions actually translate to greater reductions into ambient ozone concentrations.

We compiled a national data source linking information on power plants, ambient pollution, population demographics, and weather. The resulting data set consists of 473 power generating facilities powered by either coal or natural gas during June, July and August 2004, which represents the peak ozone season in a year following the institution of important NO_x and ozone regulations. Covariate information ($\mathbf{X}$) on each facility includes power plant operating characteristics such as operating capacity and heat input, as well as area level characteristics such as temperature and population demographics. As a measure of ozone in the area surrounding each power plant (Y), we use the fourth highest daily ozone concentration, averaged across all monitoring locations within a 100km radius. This measure is chosen to mimic the National Ambient Air Quality Standard for ozone, which is based on the annual fourth-highest daily maximum 8-hour average ozone concentration. Section 2.7.6 has a detailed description on the exact construction of the data set used in the final analysis, including references to publicly-available raw data sets and R scripts used for data construction and linkage.

We consider as "treated" the power plants for which at least 50% of the heat input is to facility units with at least one SCR or SNCR technology installed ($Z = 1$, 152 facilities), with the remaining plants regarded as untreated ($Z = 0$, 321 facilities). 67.7% of facilities have either 0% or 100% of their heat input used by units with installed SCR or SNCR control technologies, suggesting robustness to the 50% cutoff. Figure 2.1 shows maps of the power plants' treatment assignment and ozone measurements for the surrounding area, and Section 2.7.6 discusses the emission control actions of the control group $Z = 0$.

To estimate the effect of SCR/SNCR technologies relative to alternatives, we implement the "naïve" approach, Matching within Distance Caliper (distance caliper was set to 354 miles, the 15^{th} percentile of all treated-control distances), GBM, the method of Keele et al. (2015), and DAPSm (with standardized geo-distance). While the method of Keele et al. (2015) was implemented with the propensity score for the performance comparison in Section 2.4, here it is implemented in a manner more consistent with the intent of integer

programming methods, matching directly on covariates. The tolerance level for Keele et al. (2015) was set to 0.15 standard deviations for all continuous covariates. Multiple values were tried for the tuning parameter λ, and results are presented with $\lambda = 800$ (41^{st} quantile of pairwise distances) such that the number of matched pairs will be similar to that of DAPSm. The caliper used for each method was decided such that methods would balance observed covariates, where "balance" is judged by an ASDM less than 0.15 (the same value used as the tolerance for the method of Keele et al. (2015)).

The variables that are included in the propensity score model are listed in Figure 2.4 and Section 2.7.7. Characteristics of the power plants (e.g. energy consumed, compliance scheme) are not expected to exhibit strong spatial patterns, but characteristics of the surrounding areas (e.g., temperature, population demographics) are.

2.5.1 Covariate balance, number and distance of matched pairs

Covariate balance was assessed by comparing the covariate distribution of treated and control units. Without adjustment, 10 out of 18 covariates were imbalanced between the treated and control facilities, as evidenced by the leftmost values of each panel in Figure 2.4. DAPSm was performed with values of w ranging from 0 to 1, with covariate balance evaluated for each w, and depicted in the remaining portions of Figure 2.4. Note the change in covariate balance between the unadjusted setting and the setting with DAPSm($w = 0$), which matches observations based solely on proximity. Most area level characteristics achieve balance when matching only on proximity, but imbalance for power-plant level characteristics persists. Increasing values of w place more emphasis on observed propensity score differences, and balance for covariates representing power-plant characteristics improves, without a strong sacrifice in balance for the area-level covariates. Using the procedure described in Section 2.3.5, $w \approx 0.513$ was chosen for the analysis. Table 2.2 summarizes the covariate balance for all methods. GBM failed to return balanced matched samples, and is excluded from the results.

Table 2.2 also presents the number and mean distance of matched pairs. Number of matches ranged between 116 and 137, indicating that not all of 152 treated units were matched. Nonetheless, all methods should closely approximate the ATT in a manner that

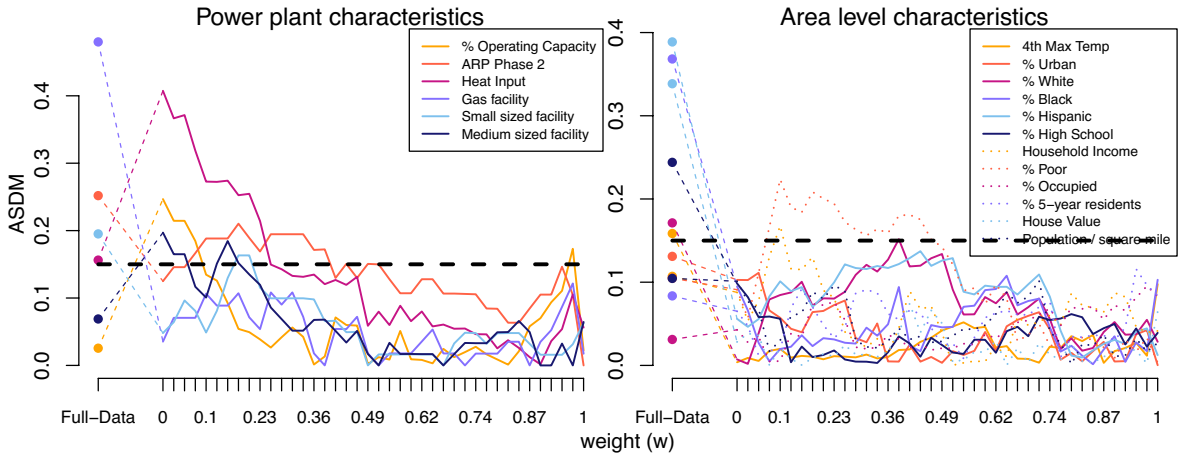


Figure 2.4: Absolute standardized difference of means for covariates that are included in the propensity score model for the full data before any matching, and for various specification of the DAPSm weight. Balance of covariates on power plant characteristics is described on the left, and balance of area-level variables is shown on the right.

Table 2.2: Balance of covariates assessed by the absolute standardized difference of means (ASDM)

	Naive	Distance Caliper	Keele et al	DAPSm	Full-data
Number of imbalanced variables	0	0	0	0	10
Mean ASDM	0.067	0.052	0.065	0.045	0.189
Max ASDM	0.148	0.134	0.145	0.150	0.480
Number of matches	137	116	124	124	
Mean distance (in miles)	1066	198	146	141	

is comparable across methods since most treated units are matched in all implementations. Characteristics of the matched population according to each method can be found in Section 2.7.8. Dropped treated units were smaller, mostly gas-operating facilities in urban areas compared to the matched treated units. Maps of the matched pairs are shown in Figure 2.5.

2.5.2 Effect estimates for NO_x and ozone

We evaluate the effectiveness of SCR/SNCR technology for reducing NO_x emissions and ambient ozone. Since emissions are measured at the power plant, the analysis of NO_x emissions is not expected to suffer from unmeasured spatial confounding. Since the formation of ambient ozone in the areas surrounding power plants is determined in part by

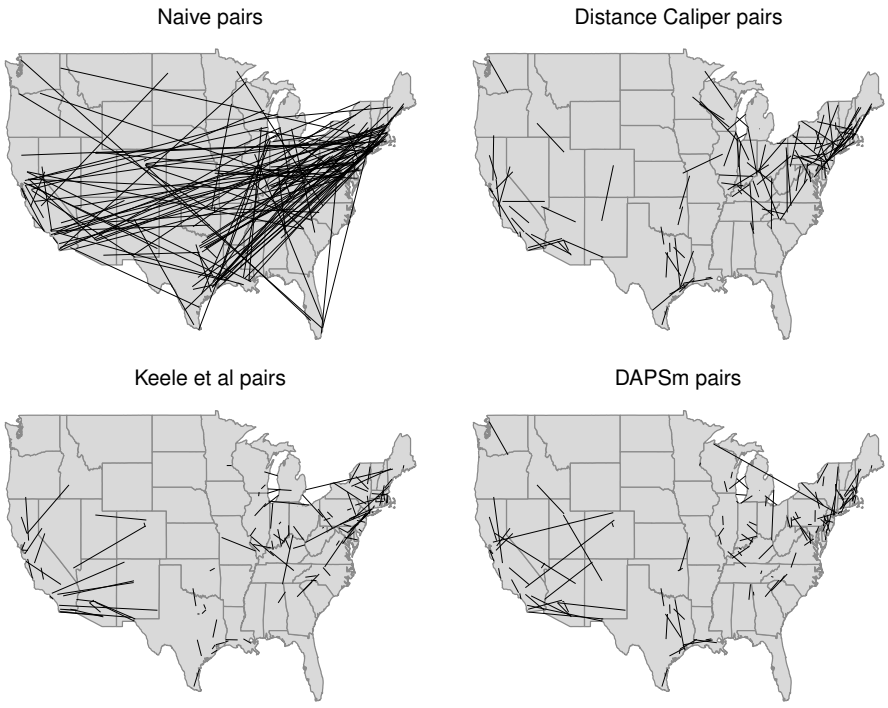


Figure 2.5: Maps of matched pairs for naïve, distance caliper, Keele et al. (2015), and DAPSm approaches. Each line segment connects one treated power plant to its matched control.

atmospheric conditions, the analysis of ozone is expected to be susceptible to unmeasured spatial confounding. Confidence intervals are constructed conditional linear models fit to the matched data sets (Ho et al., 2007). Results from all methods are reported in Table 2.3 and Figure 2.6.

Effects of SCR/SNCR on power plant NO_x emissions

Point estimates for the effect of SCR/SNCR on NO_x emissions were below zero across all methods, with the naïve and DAPSm returning significant results at the 95% confidence level. Power plants with installed SCR/SNCR emission control technologies emitted on average 205 tons of NO_x less (95% CI: 4 to 406 tons of NO_x according to DAPSm) than what they would have had emitted had they adopted an alternative NO_x control strategy.

Effects of SCR/SNCR on ambient ozone

In the analysis of ambient ozone concentrations, for which unmeasured spatial confounding is a concern, the naïve approach estimates a significant positive effect of SCR/SNCR installation on ambient ozone, which is inconsistent with the knowledge that SCR/SNCR reduces NO_x emissions and the documented relationship between NO_x and ozone. This result corroborates suspicion of unmeasured confounding. In contrast, estimates from all methods that incorporate spatial information provide estimates very close to zero (DAPSm: -0.27 parts per billion, 95% CI: -2.1 - 1.56), indicating that SCR/SNCR does not reduce ambient ozone more than alternative strategies. For reference, these effect estimates can be compared against the national ozone air quality standard of 70 parts per billion.

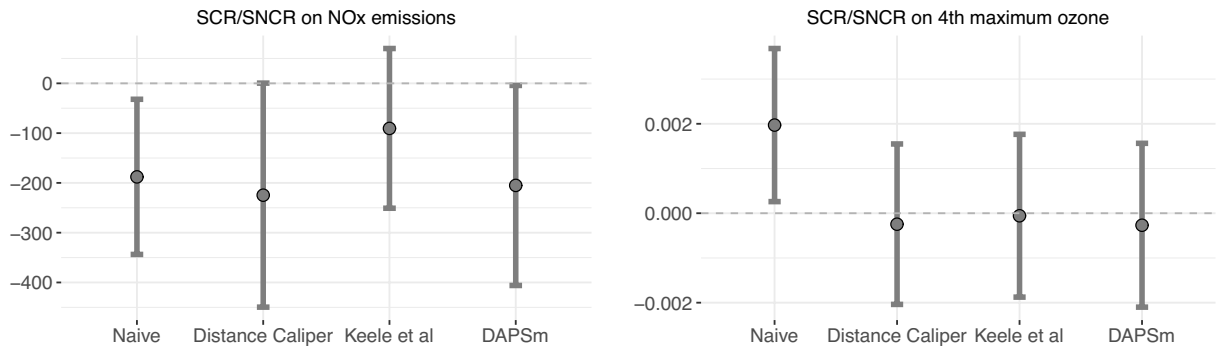


Figure 2.6: Effect estimates and 95% confidence intervals for SCR/SNCR emission control technology installation on NO_x emissions and 4^{th} maximum ozone concentration during June-August 2004, using the naïve, Matching within Distance Caliper, Keele et al. (2015), and DAPSm approaches.

Table 2.3: Estimates and 95% confidence intervals for the effect of SCR/SNCR on total NO_x emissions (in tons) and 4^{th} maximum ozone measurement (in parts per billion).

	NO_x emissions			Ozone		
	LB	Estimate	UB	LB	Estimate	UB
Naive	-343.7	-187.9	-32.0	0.26	1.97	3.68
Distance Caliper	-449.6	-224.6	0.4	-2.04	-0.24	1.55
Keele et al	-250.9	-90.5	70	-1.88	-0.06	1.76
DAPSm	**-406.1**	**-205.1**	**-4.1**	**-2.1**	**-0.27**	**1.56**

Comparison of effect estimates across methods

As mentioned earlier, since NO_x emissions are measured at the power plants' smokestacks, we do not expect unobserved spatial predictors of the outcome for this analysis. In fact, estimates across all methods are similar.

However, in the analysis of ozone concentrations, we see that the spatial methods return results that are inconsistent with the naïve method. In Section 2.7.9, we provide additional evidence of the potential of unobserved spatial confounding in the analysis of ozone, by performing a sensitivity analysis of the DAPSm effect estimates as a function of w. The sensitivity analysis corroborates the existence of an unmeasured spatial confounder, with effect estimates that increase with w (for $w > 0.513$) and approach the estimates from the naïve analysis when $w = 1$ and no adjustment for spatial proximity is made. In contrast, the sensitivity analysis of the effect of SCR/SNCR on NO_x emissions indicates that spatial confounding is not an issue.

2.6 Discussion

Unobserved confounding is a ubiquitous issue in the analysis of observational studies. Settings with spatially-indexed data provide an opportunity to recover information on unobserved spatial confounding, but most methods have been confined to regression-based approaches. We propose a method that extends the benefits of propensity score matching procedures to settings with spatially-indexed data and provide a transparent and principled framework for assessing the relative trade offs of prioritizing observed confounding adjustment and spatial proximity adjustment.

The simulation study showed the potential for DAPSm to recover information on unobserved spatial confounding. When deployed to evaluate the effectiveness of emission control technologies, DAPSm balanced all observed covariates in the resulting matched data set while providing protection against the existence of unobserved spatial confounders. The importance of incorporating spatial information was underscored by the ability of DAPSm (and other methods accounting for proximity) to return estimates that are more in line with subject-matter knowledge, in contrast to the naïve approach that ignores the

possibility of spatial confounding. Whereas the naïve approach indicated that clear reductions in NO_x were accompanied by increases in ambient ozone, analysis with DAPSm (and other methods) provided the more credible result that SCR/SNCR do not decrease ambient ozone more than other strategies.

While we compare DAPSm against Keele et al. (2015) for illustration, it is improtant to remember the important fundamental distinction between these methods: DAPSm uses the propensity score while Keele et al. (2015) propose an integer programming method that matches on covariates directly. While a comparison between propensity score methods and integer programing methods is not the goal of this paper, it is worth noting that the most salient operational difference of the two methods relates to their respective tuning parameters that govern the amount of emphasis placed on matching observations that are geographically close. DAPSm involves the tuning parameter (w) that offers a characterization of the price paid (in terms of observed covariate distance) by increasing emphasis on spatial proximity. This was evident in the ability to offer a practicable way to select a value of w (as described in Section 2.3.5), and the transparent trade off between spatial proximity and observed covariate distance is an important feature of DAPSm that aligns with a scientific goal at the forefront of air pollution (and other) studies. On the other hand, the method of Keele et al. (2015) entails a tuning parameter (λ) that balances emphasis on spatial proximity against number of obtained matches for a fixed tolerance of covariate imbalance. Keele et al. (2015) provide an approach where, for a fixed tolerance, λ could be chosen to obtain a target number of matches. Further extensions growing from the mixed integer programing literature could give rise to alternative ways of prioritizing covariate balance, the number of matches, and the relative proximity of matches.

Furthermore, we evaluated the method of Keele et al. (2015) in simulations and in comparison with other reasonable approaches, in addition to DAPSm. These simulations showed that, across a variety of spatial confounding surfaces, DAPSm with an appropriately chosen w performed comparably or better than the method of Keele et al. (2015) based on the propensity score, at least for some choices of λ.

While the comparison of different methods in the analysis of power plant emission controls highlights the potential for DAPSm to adjust for unmeasured spatial confounding,

there are several important limitations to the analysis. First, unmeasured (spatial or non-spatial) confounding may persist due to power plant or area level characteristics not contained in the data sources used. Second, we considered an "active control" group of power plants that did not install SCR/SNCR, but may have employed other strategies that could, in principle, be installed alongside SCR/SNCR (211 out of 311 control units employed a NO_x control strategy other than SCR/SNCR). Finally, the analysis relied on a simplification that linked each power plant to ambient ozone concentrations within a 100km radius. Importantly, this does not fully capture the phenomenon of long-range pollution transport whereby emissions from a particular source travel across large distances during conversion to ambient pollution. Thus, installation of control technologies at a given power plant could affect ambient pollution concentrations around power plants located at distances greater than 100km, a phenomenon referred to as "interference". While interference is not expected in the analysis of NO_x emissions, ignoring interference in the analysis of ozone concentrations has potential consequences. The simplifications used here are expected to yield estimates that are closer to zero than any true effect of SCR/SNCR on ambient ozone, as installation of these technologies is likely to reduce ambient ozone even around power plants that were considered in the "control group" for this analysis. Methods for causal inference with interference have been recently considered with spatially-indexed data (Verbitsky-Savitz and Raudenbush, 2012; Zigler et al., 2012), including our own current work on methods advances to address interference in this specific setting (Papadogeorgou et al., 2017). Furthermore, the analysis relies on some extent on correct specification of the propensity score model, and King and Nielsen (2016) argue against the use of propensity score for matching altogether. For that reason, checking covariate balance in the design phase (Rubin, 2008) without evaluating outcomes, is an important component of propensity score matching.

2.7 Appendix

2.7.1 Alternative definition of standardized distance

Another specification of the distance measure in DAPS could be the empirical CDF of all treated-control pairwise distances. Using this definition, treated unit i and control unit j have distance defined as:

$$Dist_{ij} = \frac{\sum_{k \in S_t, l \in S_c} \mathbb{I}\{d_{kl} \leq d_{ij}\}}{N_t \times N_c}. \tag{A.1}$$

With both definitions of the distance measure, $Dist_{ij} \in [0, 1]$ for every i treated, and j control.

In the simulations, the performance of DAPSm was similar with respect to MSE, absolute bias, number of matches, and balance of observed and unobserved covariates for the two specifications of standardized distance. In general, specifying distance as in (A.1) led to a smaller w chosen than the specification of (2.2). However, this is expected when examining the relationship between the two distance measures.

2.7.2 Greedy DAPSm algorithm

Consider the DAPS table of all pairs defined as:

Table 2.4: The matrix of calculated DAPS for treated-control pairs.

		Controls			
		1	2	...	N_c
Treated	1	$DAPS_{11}$	$DAPS_{12}$	...	$DAPS_{1N_c}$
	2	$DAPS_{21}$	$DAPS_{22}$	...	$DAPS_{2N_c}$
	⋮			⋮	
	N_t	$DAPS_{N_t1}$	$DAPS_{N_t2}$	...	$DAPS_{N_tN_c}$

Entries are set to infinity if the DAPS value of the pair is larger than the caliper.

The minimum element of each row is identified. These minimum values correspond to the minimum $DAPS$ for every treated unit across all controls, and rows with minimum value equal to infinity are dropped (treated units without any control within the caliper).

The matrix is reordered in increasing order of these minimum values, and the controls that acheive them are identified for every treated unit.
If there is no overlap in the control units, all treated units are matched to the controls with the smallest corresponding $DAPS$ value. Otherwise, treated units are matched up to the first control that is repeated. In that case, the new minimum $DAPS$ values are calculated over the rows of the new matrix (with matched rows and columns dropped), and the procedure is repeated.
For data sets large enough to preclude the practicality of employing DAPSm with many values of w, an algorithm which is based explicitly on assuming a non-increasing trend of ASDM with w and uses fewer fits of DAPSm can be employed. The procedure is initiated at $w = 0.5$ (step $k = 1$). If balance is achieved at step $k - 1$, w is decreased by $1/2^{k+1}$ and balance is re-accessed. If balance is not achieved at step $k - 1$, w is increased by $1/2^{k+1}$. The procedure is iterated and it stops at the value w for which the step size is smaller than a pre-specified tolerance level.

2.7.3 Data generating mechanism for simulation study

For every pair of ν, r of the spatial variable, we simulate 100 data sets. For each data set, each of 800 fixed locations are simulated to have:

1. One unmeasured confounder, U, generated as a Gaussian Process with a Matérn(ν, r) correlation function, normalized to have mean 0 and variance 1.

2. Four observed confounders, X_i, simulated as independent normal variables with mean 0, variance 1, and $Cor(U, X_i) = 0$, $i = 1, 2, 3, 4$.

3. Treatment, Z, generated as a binary variable with

$$\text{logit}P(Z = 1) = -0.85 + 0.1\, X_1 + 0.2\, X_2 - 0.1\, X_3 - 0.1\, X_4 + 0.3\, U$$

   This generative model for the treatment gives rise to data sets with approximately 30% of observations treated.

4. Outcome, Y, generated as:

$$Y = \ Z + 0.55\, X_1 + 0.21\, X_2 + 1.17\, X_3 - 0.11\, X_4 + 3\, U + \epsilon,\ \epsilon \sim N(0, 1)$$

2.7.4 Surfaces of Matérn spatial variable

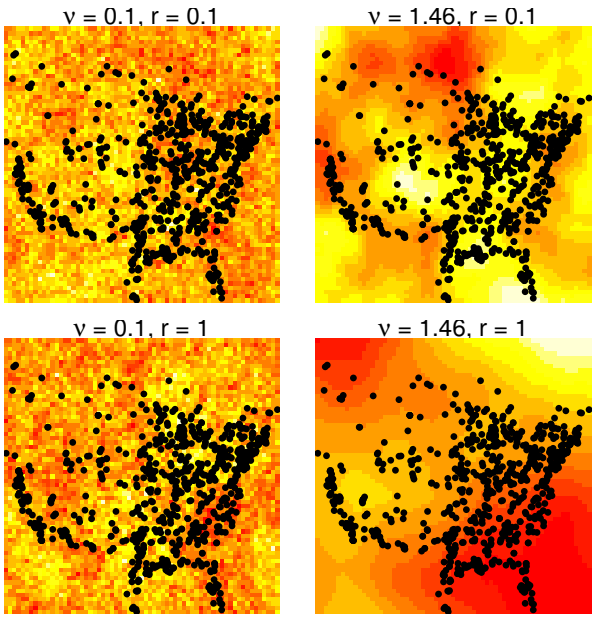


Figure 2.7: Surfaces of spatial variables with Matérn correlation function with smoothness ν and range r. Points represent 800 power plant locations. These plots correspond to the four extreme simulation scenarios.

2.7.5 Simulation results for the method of Keele et al. (2015) matching directly on covariates

In section 2.4 of the main text, we implement a simulation study to examine the performance of various methods incorporating spatial information. Most methods considered use the propensity score to adjust for observed confounders.

In the simulation study of the main text, the method of Keele et al. (2015) was implemented by matching exactly on 5 categories of the propensity score. However, Keele et al. (2015) propose their method within an integer-programming context and argue for matching directly on covariates using, for example, moment matching on continuous covariates and exact matching on discrete variables. Here, we present the relative MSE of Keele et al. (2015) for matching on covariates X_1, X_2, X_3, X_4 using moment matching with tolerance equal to 0.1 standard deviations, and a few specifications of the parameter λ.

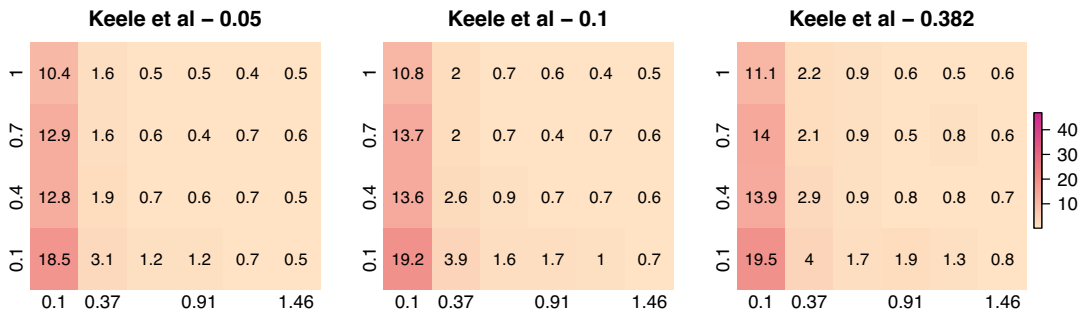


Figure 2.8: MSE of the method of Keele et al. (2015) with respect to the gold standard propensity score matching approach for values of $\lambda \in \{0.05, 0.1, 0.382\}$. Moment matching was performed for the continuous observed variables with tolerance set to 0.1 standard deviation.

The mean (IQR) number of dropped treated units for values of $\lambda \in \{0.05, 0.1, 0.382\}$ was 11.7 (8,15), 2.2 (0,3), and 0.02 (0,0) accordingly.

2.7.6 Constructing the analysis data set

All tools and data sets required to construct the analysis data set are publicly available and easily accessible. They include:

1. Raw data files (AMPD-EIA on power plants, AQS temperature data, AQS ozone data, Census 2000 data) available at `https://dataverse.harvard.edu/dataverse/dapsm`.
2. The DAPSm R Package (available at `https://github.com/gpapadog/DAPSm`)
3. The AREPA R Package (available at `https://github.com/czigler/arepa`)
4. R scripts that perform the data manipulation and linking of the raw data sets (available at `https://github.com/gpapadog/DAPSm-Analysis`).

A power plant can consist of more than one energy generating unit (EGU). We restrict our analysis to power generating units that are using coal or natural gas as one of their primary fuels. Power plant covariate information is monthly and measured for each of its EGUs. EGUs that were retired, not operating, not yet operating during June, July and

August of 2004, or did not have any data before or started having data after this time period were dropped.

A key covariate measured at the EGU level is heat input describing the energy used by the power plant unit for operation, and is therefore a good predictor of its size. Over the period of three months (June, July, August 2004), 14% of all entries lack heat input information. 82% of the missing data was imputed using heat input information from other years (2003, 2002, 2005, 2006) (R squared of the models used ranges between 89.8-94.7%). 25% of predicted heat input observations were estimated to close to zero but negative. These values were set to 0. Monthly unit level data were afterwards aggregated to the facility level and over the June-August 2004 period. 68 (5%) facilities were dropped because of missing heat input.

For each ozone monitoring site, temperature information was assigned as the average temperature over all temperature monitoring sites within 150 kilometers. Population demographics for the year 2000 for the surrounding area were assigned to each ozone monitor as aggregated zip code-level Census variables with centroids located within 6 miles of the monitor. The resulting data set includes ozone, temperature and demographics measured at or around ozone monitoring sites.

Finally, ozone, temperature and population demographics were assigned to power plants as the average over ozone monitors within 100 kilometers. Each monitor was only allowed to contribute to the closest power plant. A monitor site is not linked to any power plant if there is no power plant within 100 kilometers. A power plant is not linked to any ozone monitor, if there is no ozone monitor within 100 kilometers without another power plant closer.

The resulting data set consists of 483 power plants linked to a total of 937 ozone monitors. 10 additional facilities are dropped due to missing Census information, or missing percent capacity, resulting to 473 facilities in our final data set linked to a total of 921 ozone monitoring sites.

A facility is considered treated if at least 50% of its heat input is used by EGUs with at least one SCR/SNCR installed. 1,230 out of 2,964 EGUs in control facilities have no NO_x emission control technologies installed, while 33 have one of SCR/SNCR installed

(amounting to less than 50% of the facility's heat input). The remaining units that constitute the facilities in the control group have some other type of NO_x control installed such as a low NO_x burner, an overfire reduction system, an ammonia injection system, a modified combustion method, water injection system, or some other non-specified control. All of these alternative strategies in the control group are designed to reduce NO_x, but are widely regarded as less efficient than SCR/SNCR. Note that it is very common for EGUs (treated or control) to follow more than one NO_x emission control strategies, implying that even EGUs in the control group having other NO_x control strategies might still be candidates for additional installation of SCR/SNCR.

2.7.7 Data application covariate description

Table 2.5: Power plant and area level characteristics before matching

Name	Description	Units	Treated	Control	ASDM
% Capacity	Percentage of operating capacity	-	0.42 (0.28)	0.42 (0.3)	0.026
Heat Input	Amount of fuel energy burned for power generation (logarithm)	log MMBtu	14.3 (1.97)	14 (1.96)	0.156
Phase 2	ARP Phase 2 indicator	-	0.86 (0.35)	0.77 (0.42)	0.252
Gas	Mostly gas burning power plant	-	0.77 (0.42)	0.57 (0.5)	0.48
Small sized	Power plants consisted by 1 or 2 EGUs	-	0.62 (0.49)	0.52 (0.5)	0.195
Medium sized	Power plants consisted by 3, 4, or 5 EGUs	-	0.33 (0.47)	0.36 (0.48)	-0.069
Temperature	4^{th} maximum temperature over study period	Fahrenheit	70.7 (8.1)	69.5 (7.7)	0.158
% Urban	Percentage of population in urban areas	-	0.76 (0.32)	0.72 (0.34)	0.131
% White	Percentage of white population	-	0.77 (0.17)	0.80 (0.17)	-0.171
% Black	Percentage of black population	-	0.08 (0.09)	0.11 (0.14)	-0.368
% Hispanic	Percentage of hispanic population	-	0.16 (0.18)	0.09 (0.14)	0.389
% High School	Percentage of population that attended high school	-	0.29 (0.08)	0.31 (0.08)	-0.244
Household Income	Median household income	USD	44,721 (12,456)	43,386 (12,223)	0.107
% Poor	Percentage of impoverished population	-	0.13 (0.06)	0.12 (0.06)	0.105
% Occupied	Percentage of occupied population	-	0.91 (0.11)	0.91 (0.1)	0.031
% MovedIn5	Percentage of population that has lived in the area for less than 5 years	-	0.48 (0.08)	0.47 (0.09)	0.083
House value	Median house value	USD	148,394 (97,144)	115,510 (57,472)	0.339
Population density	Population per square mile (logarithm)	log # / mile^2	6.19 (1.68)	6.02 (1.71)	0.105

2.7.8 Description of matched and dropped treated units

In Table 2.6 we show the mean and standard deviation of the matched and dropped treated units for each method. The mean and standard deviation of the treated units in the full data can be found in Table 2.5.

Table 2.6: Covariate mean and standard deviation for matched and dropped treated units for the Naïve, Matching in Distance Caliper, Keele et al. (2015), and DAPSm.

	Naïve	Distance Caliper	Keele et al. (2015)	DAPSm
Matched units				
% Capacity	0.43 (0.29)	0.4 (0.27)	0.36 (0.27)	0.42 (0.28)
Heat Input	14.4 (1.96)	14.1 (1.98)	13.8 (2.07)	14.3 (2.05)
Phase 2	0.85 (0.36)	0.86 (0.35)	0.91 (0.29)	0.82 (0.38)
Gas	0.69 (0.46)	0.81 (0.39)	0.85 (0.36)	0.73 (0.45)
Small sized	0.58 (0.5)	0.72 (0.45)	0.70 (0.46)	0.59 (0.49)
Medium sized	0.36 (0.48)	0.25 (0.43)	0.26 (0.44)	0.36 (0.48)
Temperature	70.1 (7.59)	70.8 (7.79)	71.5 (7.72)	70.6 (8.59)
% Urban	0.79 (0.29)	0.78 (0.31)	0.79 (0.3)	0.72 (0.34)
% White	0.77 (0.17)	0.76 (0.17)	0.75 (0.18)	0.79 (0.16)
% Black	0.08 (0.1)	0.06 (0.07)	0.06 (0.06)	0.08 (0.1)
% Hispanic	0.15 (0.19)	0.18 (0.19)	0.20 (0.21)	0.13 (0.17)
% High School	0.30 (0.08)	0.28 (0.08)	0.28 (0.08)	0.31 (0.08)
Household Income	44166 (10964)	45430 (11994)	45846 (11748)	44496 (12654)
% Poor	0.12 (0.06)	0.13 (0.06)	0.13 (0.06)	0.12 (0.06)
% Occupied	0.92 (0.05)	0.93 (0.05)	0.92 (0.06)	0.90 (0.12)
% MovedIn5	0.48 (0.07)	0.49 (0.07)	0.49 (0.07)	0.47 (0.08)
House value	142750 (79501)	155577 (83956)	161888 (86553)	137894 (98511)
Dropped units				
% Capacity	0.42 (0.27)	0.44 (0.28)	0.47 (0.28)	0.44 (0.25)
Heat Input	14.25 (2)	14.51 (1.96)	14.69 (1.83)	14.25 (1.63)
Phase 2	0.87 (0.34)	0.86 (0.35)	0.81 (0.39)	1 (0)
Gas	0.87 (0.34)	0.73 (0.44)	0.71 (0.46)	0.96 (0.19)
Small sized	0.67 (0.47)	0.53 (0.5)	0.56 (0.5)	0.75 (0.44)
Medium sized	0.28 (0.45)	0.4 (0.49)	0.38 (0.49)	0.18 (0.39)
Temperature	71.5 (8.7)	70.7 (8.39)	70.19 (8.38)	71.52 (5.46)
% Urban	0.72 (0.35)	0.75 (0.32)	0.74 (0.33)	0.92 (0.11)
% White	0.77 (0.17)	0.78 (0.17)	0.78 (0.16)	0.65 (0.16)
% Black	0.07 (0.09)	0.09 (0.11)	0.1 (0.11)	0.06 (0.07)
% Hispanic	0.17 (0.17)	0.14 (0.18)	0.13 (0.16)	0.29 (0.17)
% High School	0.28 (0.08)	0.3 (0.08)	0.31 (0.08)	0.23 (0.04)
Household Income	45425 (14180)	44131 (12871)	43858 (12976)	45719 (11702)
% Poor	0.13 (0.06)	0.13 (0.06)	0.13 (0.06)	0.14 (0.05)
% Occupied	0.89 (0.15)	0.9 (0.14)	0.9 (0.13)	0.95 (0.02)
% MovedIn5	0.48 (0.08)	0.48 (0.08)	0.47 (0.08)	0.53 (0.04)
House value	155557 (115989)	142424 (107020)	138040 (103856)	194898 (76283)

2.7.9 DAPSm effect estimates as a function of the tuning parameter

We investigate the sensitivity of the DAPSm results to the specification of w, by plotting the estimates and 95% confidence intervals of the "effect" estimates as a function of w. First, note the number of imbalanced covariates is generally decreasing as a function of w (Figure 2.9), indicating that a weight over the chosen 0.513 is necessary to balance

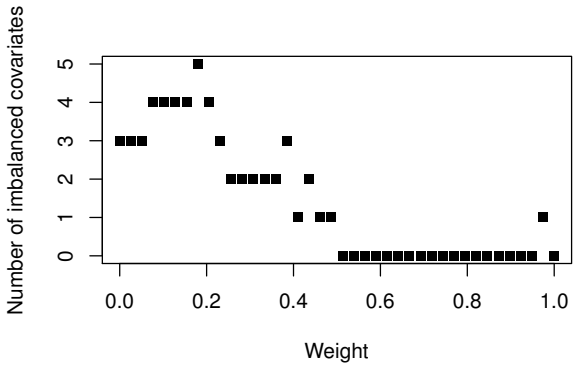


Figure 2.9: Number of imbalanced variables as a function of w in DAPSm.

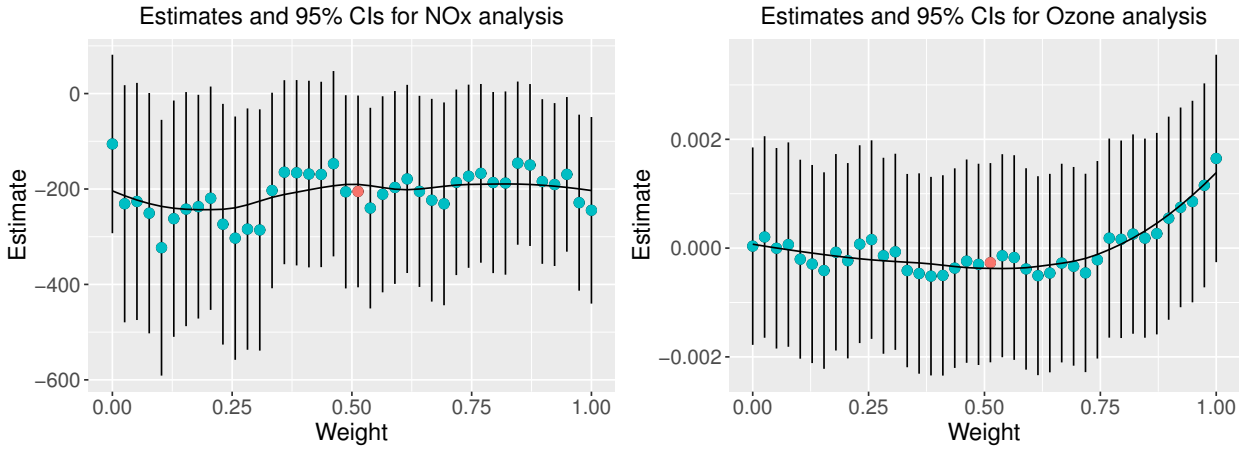


Figure 2.10: Estimates and 95% confidence intervals using DAPSm as a function of w for the NO_x emissions (left) and O_3 (right) analyses. The red point corresponds to the effect estimate based on $w = 0.513$.

all observed covariates. Therefore, the estimates for small values of w are not necessarily interpretable, since residual confounding might still be present. As the value of w increase above the chosen value, balance of observed covariates is almost always maintained, while less weight is given to achieving matches at close proximity. As w tends to 1, DAPSm resembles simple propensity score matching and incorporates a decreased amount of spatial information.

As mentioned on the main text, unobserved spatial confounding is unlikely to be present in the analysis of NO_x emissions, since NO_x emissions are measured directly at the plant's smokestacks, and area-level characteristics are unlikely to be predictors. In the left panel

of Figure 2.10, we see that the effect estimates for the NO_x analysis remain more or less constant for values of $w \geq 0.513$, indicating that as long as observed covariates are balanced, the effect estimates will be similar and independent of the chosen value of w. However, in the right panel of Figure 2.10, we see a very different result. Specifically, the effect estimates are increasing for values of w greater than 0.513. This might imply that, when observed covariates are balanced, matching on spatial proximity is an important component of acquiring unbiased effect estimates. Specifically, it indicates that there might be an unobserved spatial confounder, which leads to positive bias when it is not adjusted for.

Causal inference with interfering units for cluster and population level treatment allocation programs

Georgia Papadogeorgou

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Fabrizia Mealli

Department of Statistics, Computer Science, Applications

University of Florence

Corwin M. Zigler

Department of Biostatistics

Harvard Chan School of Public Health

3.1 Introduction

Most causal inference literature assumes that a unit's potential outcome depends solely on its treatment, and does not depend on the treatments of other units in the population. However, this assumption is often not reasonable. Perhaps the most classical example arises in vaccination studies (Ali et al., 2005; Hudgens and Halloran, 2008) where a unit's disease status depends on their own vaccination status but also on the vaccination status of others in their social network. The presence of interference can lead to misleading results for familiar causal estimands (Sobel, 2006), or estimands that lack clear causal interpretation (Tchetgen Tchetgen and VanderWeele, 2012), but can also introduce new estimands of intrinsic scientific interest.

Sobel (2006) defined estimands for interference when the population can be partitioned into clusters for which a unit's potential outcomes depend only on the treatment of units within the same cluster. Such assumption is called *partial interference*, and the interference clusters are also called interference groups. Hudgens and Halloran (2008) formalized causal inference in the presence of interference in the context of two-stage randomization designs, which was extended to observational studies by Tchetgen Tchetgen and VanderWeele (2012), and Perez-Heydrich et al. (2015).

In order to continue development in the context of observational studies, we highlight a key distinction that arises when formulating average potential outcomes in the presence of interference, which generally requires consideration of vectors of treatment assignments. We use the term *treatment allocation strategy* to refer to a process giving rise to either observed or hypothesized vectors of treatment assignments. The *observed treatment allocation strategy* refers to that which gives rise to observed treatments. The *counterfactual treatment allocation strategy* refers to how treatments may have been assigned in some hypothesized counterfactual world for which causal contrasts can be considered. This distinction between observed and counterfactual treatment allocation programs helps illuminate that existing causal estimands, such as those in Tchetgen Tchetgen and VanderWeele (2012), are limited to counterfactual treatment allocation programs that remain agnostic with regard to covariate information (as would be the case in a two-stage random-

ized study). These estimands ignore the possible role of unit-level covariates that relate to treatment adoption, implicitly assuming an intervention manipulating each individual unit's treatment propensity. Consequently, these estimands pertain to counterfactual worlds where, for example, treatments are allocated to units according to a Bernoulli distribution with equal probability for each unit within a cluster.

In many settings, however, treatment allocations corresponding to unit-level manipulation are difficult to conceive. For example, policy interventions may be designed to increase the prevalence of a treatment without direct control over the individual treatment propensity. In such settings, individual treatment adoption might generally depend on unit-level covariates or the treatment status of neighboring units. To address such settings, we develop new causal estimands anchored to counterfactual treatment allocations that are conceived *at the cluster level*, where a particular allocation strategy dictates the cluster-average propensity of receiving treatment without directly specifying individual-level treatment propensities. Specifically, under the assumption of partial interference, we introduce estimands for counterfactual treatment allocation programs which 1) do not assume unit-level manipulation of treatment propensities; 2) but allow for correlation of treatment assignment within a cluster; and 3) unit-level propensity of treatment that depends on individual and group level covariates. Note that, in focusing on new estimands for covariate-dependent counterfactual treatment allocations programs, our work has commonalities with independent ongoing work in Barkley et al. (2017).

Causal inference may also be motivated to investigate interventions on the distribution of cluster-average treatment propensities. Such may be the case when evaluating policies that are not designed to manipulate individual or cluster-average treatment propensity, but rather change the distribution of cluster-average propensities of receiving treatment by, for example, providing a population-wide incentive to adopt treatment. Accordingly, we also define estimands for counterfactual treatment allocation strategies defined *at the population level* that shift the distribution of the cluster-average propensity of receiving the treatment, without specifying the average treatment propensity of any specific cluster.

Definition of the new causal estimands described above is accompanied here by new estimators and derivation of corresponding asymptotic properties as the number of clus-

ters grows. Related work can be found in Ferracci et al. (2014). Other relevant work includes Liu and Hudgens (2014) where asymptotic results are derived for growing number of clusters or number of individuals within clusters, Perez-Heydrich et al. (2015) where large sample variance estimators for the estimator of Tchetgen Tchetgen and VanderWeele (2012) are derived, and Liu et al. (2016), where estimands and estimators are extended to the case of a network where partial interference does not hold, but asymptotic results are derived under the assumption of partial interference.

The motivating context for this work is the evaluation of interventions to limit harmful pollution from power plants that are geographically clustered. The movement of air pollution through space leads to interference: intervening on one power plant can affect the air pollution surrounding nearby power plants. Existing estimands such as those in Tchetgen Tchetgen and VanderWeele (2012) represent quantities for counterfactual treatment allocations in two steps where 1) a constant treatment probability governs the proportion of power plants that would be "treated" within a cluster, and 2) based on that probability, power plants within the cluster are randomly and independently assigned the treatment. However, this structure does not cohere to that of air pollution regulations, where, in reality, the adoption of treatments at power plants is not directly mandated and is heavily influenced by power-plant characteristics (e.g., the size or operating capacity of the plant). Instead, regulatory programs often work by incentivizing regions of power plants to adopt certain technologies (e.g., by changing the penalties for over-emission), but which power plants actually adopt them is highly dependent on covariates and may be spatially correlated. Thus, new estimands for counterfactual treatment allocations where individual-level treatment adoption depends on covariates - subject to a cluster average treatment propensity - coheres more closely to the realities of air pollution regulations. Additionally, estimands at the population level of clusters could refer to counterfactual situations where some higher level of government (e.g., federal) issues additional incentives for power plants to install the technologies, but cannot mandate installation, and different regions can comply to different degrees. The new estimators are deployed here to an analysis of U.S. power plants investigating the comparative effectiveness of Selective Catalytic or non-Catalyitic Reduction systems (relative to other

strategies) for reducing ambient ozone pollution. A preliminary investigation of these same data in Papadogeorgou et al. (2018) ignored interference and indicated that these systems causally reduced NO_x emissions (an important precursor to ozone pollution) but did not lead to a reduction in ambient ozone. The analysis here to address the possibility of interference produces meaningfully different results that are more consistent with the literature relating NO_x emissions to ambient ozone pollution. Note that, despite the focus on air pollution interventions, similar considerations could be construed in more classical interference settings such as vaccine studies, where certain types of community members might be more likely to receive the vaccine and vaccine programs may be designed to increase vaccine coverage at the community, or national level.

In Section 3.2 we introduce the notation and the new estimands for the cluster-level intervention, for which estimators are presented in Section 3.3, along with unbiasedness, consistency and asymptotic distribution results for an increasing number of clusters. In Section 3.4, the estimand for interventions at the population of clusters level is introduced. The rest of the paper presents some simulations in Section 3.5, our data application in Section 3.6 and concludes with some discussion on the limitations and future directions of this paper in Section 3.7.

3.2 Estimands under partial interference

We adopt the notation used in Tchetgen Tchetgen and VanderWeele (2012). Let N be the number of clusters, and n_i the number of units in cluster i, $i \in \{1, 2, \ldots, N\}$. Furthermore, denote $\boldsymbol{A}_i = (A_{i1}, A_{i2}, \ldots, A_{in_i}) \in \mathcal{A}(n_i)$ to be the cluster treatment vector, and $\boldsymbol{A}_{i,-j} = (A_{i1}, A_{i2}, \ldots, A_{ij-1}, A_{ij+1}, \ldots, A_{in_i}) \in \mathcal{A}(n_i - 1)$ to be the treatment of all units in cluster i apart from unit j, where $\mathcal{A}(n) = \{0, 1\}^n$. Furthermore, let L_{ij} be a vector of individual and cluster-level covariates, and $\boldsymbol{L}_i = (L_{i1}, L_{i2}, \ldots, L_{in_i})$ be the collection of covariates of all units within a cluster.

Under the assumption of partial interference, the potential outcome of unit j in cluster i may depend on the treatment of units in cluster i, but not on the treatment of units in different clusters. For every i we postulate the existence of group i's potential outcomes

$\boldsymbol{Y}_i(\cdot) = \{\boldsymbol{Y}_i(\boldsymbol{a}_i), \boldsymbol{a}_i \in \mathcal{A}(n_i)\}$, where $\boldsymbol{Y}_i(\boldsymbol{a}_i) = (Y_{i1}(\boldsymbol{a}_i), Y_{i2}(\boldsymbol{a}_i), \ldots, Y_{in_i}(\boldsymbol{a}_i))$.

3.2.1 Average potential outcome

Under the assumption of partial interference, we define the individual average potential outcome for a counterfactual treatment allocation strategy with two features: 1) treatment assignment for units within a cluster is unlikely to be independent, and 2) individual covariates can be predictive of a unit's treatment probability. Let $P_{\alpha,L}$ represent the (arbitrarily specified) counterfactual treatment allocation program, specified specifically to depend on covariates and/or allow correlated assignments within clusters. $P_{\alpha,L}$ is governed by parameters α, which represent features of the counterfactual treatment allocation program of interest. For the purpose of this paper, we consider α to represent the cluster-average propensity of treatment.

The individual average potential outcome is defined as:

$$\overline{Y}_{ij}^{L}(a;\alpha) = \sum_{\boldsymbol{s}\in\mathcal{A}(n_i-1)} Y_{ij}(A_{ij}=a, \boldsymbol{A}_{i,-j}=\boldsymbol{s}) P_{\alpha,L}(\boldsymbol{A}_{i,-j}=\boldsymbol{s}|A_{ij}=a, \boldsymbol{L}_i), \tag{3.1}$$

and represents the expected outcome for unit j in cluster i in the counterfactual world where treatment is assigned with respect to $P_{\alpha,L}$, but the treatment of unit j is fixed to a. This estimand is well-defined for any fixed choice of $P_{\alpha,L}$. Based on the individual average potential outcome, group and population average potential outcomes are defined as

$$\overline{Y}_{i}^{L}(a;\alpha) = \frac{1}{n_i}\sum_{j=1}^{n_i} \overline{Y}_{ij}^{L}(a;\alpha), \tag{3.2}$$

$$\overline{Y}^{L}(a;\alpha) = \frac{1}{N}\sum_{i=1}^{N} \overline{Y}_{i}^{L}(a;\alpha). \tag{3.3}$$

3.2.2 The counterfactual treatment allocation in existing literature

As mentioned previously, $P_{\alpha,L}$ can be arbitrarily chosen and represents the process with which treatment is assigned in the counterfactual world, driving the interpretation of all estimands. The above development has left unspecified the term $P_{\alpha,L}$ in (3.1) providing

relative weights to different cluster treatment vectors in the individual average potential outcomes. The estimands in Tchetgen Tchetgen and VanderWeele (2012) and Perez-Heydrich et al. (2015) correspond to counterfactual treatment strategies $P_{\alpha,L}(\boldsymbol{a}_i|\boldsymbol{L}_i) = \pi(\boldsymbol{a}_i;\alpha) = \prod_{j=1}^{n_i} \alpha^{a_{ij}}(1-\alpha)^{1-a_{ij}}$, giving equal probability to all cluster-treatment vectors with the same number of treated units, irrespective of which those units are. For this choice of $P_{\alpha,L}$ the estimands represent quantities in counterfactual worlds where individual treatment probability can be manipulated and units are assigned to treatment independently and with equal probability α.

3.2.3 Realistic counterfactual treatment allocation program

However, in some situations, counterfactual treatment allocations can only be realistically conceived if allowed to depend on covariates or if they incorporate correlation between treatment of units in the same cluster. In the study of power plant interventions on ambient air quality, the decision of whether to "treat" a power plant is at the discretion of the power company and heavily influenced by power plant covariates. Therefore, a hypothesized counterfactual treatment allocation is realistic only when such covariates are incorporated.

As an example, consider the power-plant level covariate 'heat input', a proxy for the size of the power plant, and let L_{ij} be the heat input of power plant j in cluster i. Then, one specification of a counterfactual treatment allocation strategy that would acknowledge that different-sized power plants are more or less likely to adopt treatment is:

$$\text{logit}P_{\alpha,L}(A_{ij} = 1|L_{ij}) = \xi_i^\alpha + \delta_L L_{ij},$$

for some *fixed*, pre-specified value of δ_L, and ξ_i^α such that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} \text{expit}\left(\xi_i^\alpha + \delta_L L_{ij}\right) = \alpha.$$

The value δ_L here could be specified according to knowledge of how the size of the power plant is expected to impact the propensity to adopt treatment. Based on this specification of $P_{\alpha,L}$, the estimands of interest correspond to quantities in a hypothesized world where

treatment is assigned independently across units with treatment propensity that depends on L_{ij}, but is on average equal to α. A data-driven way to choose $P_{\alpha,L}$ is presented in Section 3.6.

This fully specifies the probability of the cluster treatment vector under the counterfactual treatment allocation $P_{\alpha,L}(\boldsymbol{A}_i = \boldsymbol{a}_i|\boldsymbol{L}_i)$ for all $\boldsymbol{a}_i$, and therefore specifies $P_{\alpha,L}(\boldsymbol{A}_{i,-j} = \boldsymbol{s}|A_{ij} = a, \boldsymbol{L}_i)$ for all $\boldsymbol{s} \in \mathcal{A}(n_i - 1)$ giving relative weights in the specification of the individual average potential outcome (3.1).

Alternatively, a counterfactual treatment allocation strategy can also be defined to incorporate dependence of treatments in the same cluster. For example, consider

$$\text{logit}P_{\alpha,L}(A_{ij} = 1|L_{ij}) = \xi_i^\alpha + \delta_L L_{ij} + \theta_{ij},$$

where θ_{ij} is a mean 0 spatial random effect with fixed correlation matrix decaying with distance. This choice of $P_{\alpha,L}$ corresponds to a counterfactual treatment allocation program that depends on covariates and incorporates dependent treatment assignment of units within a cluster.

3.2.4 Direct and indirect effects

Different contrasts of average potential outcomes can be considered to characterize how treatment affects the outcome of interest. For counterfactual allocation strategy $P_{\alpha,L}$, direct effects represent contrasts in average potential outcomes when only the individual treatment changes. On the other hand, indirect effects contrast average potential outcomes for a fixed level of individual treatment, but different specification of the parameter α governing the counterfactual allocation program. For that reason, indirect effects represent expected changes in potetntial outcomes for changes only in the "treatment of neighbors", and they can be thought of as a measure of interference. Indirect effects are also known in the literature as spillover effects.

Based on the individual, group and population average potential outcomes, one can define the individual, group and population direct effects as

$$DE_{ij}^L(\alpha) = \overline{Y}_{ij}^L(1;\alpha) - \overline{Y}_{ij}^L(0;\alpha),$$

$$DE_i^L(\alpha) = \overline{Y}_i^L(1,\alpha) - \overline{Y}_i^L(0;\alpha) = \frac{1}{n_i}\sum_{j=1}^{n_i} DE_{ij}^L(\alpha)$$

$$DE^L(\alpha) = \overline{Y}^L(1,\alpha) - \overline{Y}^L(0;\alpha) = \frac{1}{N}\sum_{i=1}^{N} DE_i^L(\alpha)$$

accordingly. Similarly, the individual indirect effect is defined as

$$IE_{ij}^L(\alpha_1,\alpha_2) = \overline{Y}_{ij}^L(0,\alpha_2) - \overline{Y}_{ij}^L(0,\alpha_1),$$

based on which group and population indirect effects can be defined. Indirect effects could be alternatively defined for individual treatment assignment $a = 1$, but here our focus is on the effect of neighbors' treatment in the areas surrounding untreated power plants. Contrasts other than the difference can also be considered. Based on these estimands, total effects can be defined as the sum of direct and indirect effects (Hudgens and Halloran, 2008), while similar development can lead to the definition of overall effects.

3.3 Estimating the population average potential outcome

For a fixed choice of $P_{\alpha,L}$, we provide estimators of the population average potential outcome in (3.3), unbiasedness and consistency results, and derive the estimator's asymptotic distribution when the number of clusters increases to infinity, for a known or correctly specified parametric cluster-propensity score model (defined below). Based on these, estimators and asymptotic distributions for the estimands in Section 3.2.4 can be acquired as demonstrated in Example 1. Proofs are in Section 3.8.3.

We start by making the sample cluster-level *positivity*, and *ignorability* assumptions:

Assumption 1. *Positivity. For $i \in \{1,2,\ldots,N\}$, the probability of observing cluster treatment vector $\boldsymbol{a}_i$ given cluster covariates $\boldsymbol{L}_i$ is denoted by $f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i = \boldsymbol{a}_i|\boldsymbol{L}_i)$ and is positive for all $\boldsymbol{a}_i \in \mathcal{A}(n_i)$. $f_{\boldsymbol{A}|\boldsymbol{L},i}$ is the cluster-propensity score.*

Assumption 2. *Ignorabililty. For $i \in \{1,2,\ldots,N\}$, the observed cluster treatment $\boldsymbol{A}_i$ is conditionally independent of the set of cluster potential outcomes $\boldsymbol{Y}_i(\cdot)$ given the covariates $\boldsymbol{L}_i$, denoted as $\boldsymbol{A}_i \amalg \boldsymbol{Y}_i(\cdot)|\boldsymbol{L}_i$.*

3.3.1 Estimators of the group and population average potential outcome

Let

$$\widehat{Y}_i^L(a;\alpha) = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} \tag{3.4}$$

$$\widehat{Y}^L(a;\alpha) = \frac{1}{N}\sum_{i=1}^{N} \widehat{Y}_i^L(a;\alpha) \tag{3.5}$$

where $f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)$ is the cluster-level propensity score for the observed treatment, and $P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)$ is the probability of the observed treatment on units other than j given $A_{ij}=a$, under the counterfactual treatment allocation program.

Assuming that the group level propensity score $f_{\boldsymbol{A}|\boldsymbol{L},i}(\cdot|\boldsymbol{L}_i)$ is known, and Assumptions 1, 2 hold, then $\widehat{Y}_i^L(a;\alpha)$, $\widehat{Y}^L(a;\alpha)$ are unbiased for $\overline{Y}_i^L(a,\alpha), \overline{Y}^L(a,\alpha)$ accordingly. Unbiasedness is derived for a fixed set of clusters with respect to the distribution of the observed treatment assignment.

The population average potential outcome (3.3) is defined as the average of the group average potential outcomes. Alternative definitions could weigh each cluster by cluster sample size (which is what the population average potential outcome of Liu et al. (2016) simplify to under the assumption of partial interference). In Section 3.8.5, we discuss this distinction and provide an argument why an equal-weight estimand and the corresponding estimator (3.5) might be preferable.

3.3.2 Asymptotic results for $\widehat{Y}^L(a;\alpha)$ for known propensity score

We derive the asymptotic properties of the estimator in (3.5) for an increasing number of clusters N, denoted by $\widehat{Y}_N^L(a;\alpha)$. Let $\widehat{Y}_N^L(\alpha) = \left(\widehat{Y}_N^L(0;\alpha),\ \widehat{Y}_N^L(1;\alpha)\right)^T$.

Assume that the N clusters are a sample of an infinite superpopulation of clusters from which they are sampled randomly. Therefore $(\boldsymbol{Y}_i(\cdot), \boldsymbol{A}_i, \boldsymbol{L}_i)$ are now independent and identically distributed random vectors, whose distribution is denoted as F_0. (For notational simplicity, n_i is included in $\boldsymbol{L}_i$.) The sample positivity and ignorability assumptions are translated to their super-population counterparts.

Assumption 3. *Super-population positivity. There exists* $\exists \rho > 0$ *such that* $f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i) > \rho$ *with probability 1.*

Assumption 4. *Super-population ignorability. For* F_0, $\boldsymbol{A}_i \amalg \boldsymbol{Y}_i(\cdot)|\boldsymbol{L}_i$.

Theorem 1. *Let* $\boldsymbol{\mu}_0 = (\mu_0^0, \mu_0^1)^T$ *where* $\mu_0^a = E_{F_0}\left[\overline{Y}_i^L(a;\alpha)\right]$. *Under Assumptions 3, 4, for known propensity score, and bounded outcome (there exists* $M > 0 : |Y_{ij}| < M$ *with probability 1),* $\widehat{Y}_N^L(\alpha)$ *is consistent for* $\boldsymbol{\mu}_0$ *and asymptotically normal with limiting distribution* $\sqrt{N}\left(\widehat{Y}_N^L(\alpha) - \boldsymbol{\mu}_0\right) \xrightarrow{d} N(0, V(\boldsymbol{\mu}_0))$, *where*

$$
\begin{aligned}
V(\boldsymbol{\mu}_0) &= E_{F_0}\left[\psi(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\mu}_0)\psi(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\mu}_0)^T\right], \\
\psi(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\mu}_0) &= \left(\psi_{0,\alpha}\left(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \mu_0^0\right),\ \psi_{1,\alpha}\left(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \mu_0^1\right)\right)^T \\
\psi_{a,\alpha}(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \mu_0^a) &= \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij} = a)Y_{ij} - \mu_0^a.
\end{aligned}
$$

The above theorem leads to the approximation $\widehat{Y}_N^L(\alpha) \overset{approx}{\sim} MVN_2\left(\boldsymbol{\mu}_0, N^{-1}V(\boldsymbol{\mu}_0)\right)$ for large number of clusters. Even if assumptions about F_0 are made, the elements of $V(\boldsymbol{\mu}_0) = Cov_{F_0}\left[\left(\overline{Y}_i(0,\alpha), \overline{Y}_i(1,\alpha)\right)^T\right]$ (proof in Section 3.8.4) are often hard to calculate analytically. Instead, the asymptotic variance of $\widehat{Y}_N^L(\alpha)$ can be estimated using the empirical expectation

$$
\widehat{V}\left(\boldsymbol{\mu}\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\psi(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\mu})\psi(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\mu})^T\right],
$$

evaluated at $\boldsymbol{\mu} = \widehat{Y}_N^L(\alpha)$. Under regularity conditions, discussed in Iverson and Randles (1989), $\widehat{V}\left(\widehat{Y}_N^L(\alpha)\right)$ will be consistent for $V(\boldsymbol{\mu}_0)$. Using Theorem 1 one can acquire the asymptotic distribution of a contrast between $\widehat{Y}^L(0;\alpha)$, $\widehat{Y}^L(1;\alpha)$ specifying a direct effect, by an application of the multivariate delta method, as shown in Example 1.

3.3.3 Asymptotic results for $\widehat{Y}^L(a;\alpha)$ for estimated propensity score from a correctly-specified parametric model

However, most of the times the propensity score is not known, and has to be estimated using the observed data. In the next theorem, we provide the asymptotic distribution of $\widehat{Y}_N^L(\alpha)$ when the propensity score is estimated using a correctly specified parametric

propensity score model. In this case, the cluster-propensity score for the observed treatment vector will be denoted by $f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i;\boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ are the model parameters.

Theorem 2. *Assume that assumptions 3, 4 hold, the outcome is bounded with probability 1 (as in Theorem 1) and the parametric form of the propensity score model indexed by $\boldsymbol{\gamma}$, $f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{a}_i|\boldsymbol{l}_i;\boldsymbol{\gamma})$, is correctly specified and differentiable with respect to $\boldsymbol{\gamma}$. Let $\boldsymbol{\mu}_0$ be as in Theorem 1, and $\widehat{Y}_N^L(a,\alpha)$ calculated using consistent estimates $\widehat{\boldsymbol{\gamma}}$ of the propensity score $f_{\boldsymbol{A}|\boldsymbol{L},i}$. Let $\boldsymbol{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}^T}\log f(\boldsymbol{a}_i|\boldsymbol{l}_i;\boldsymbol{\gamma})$ be the score functions. Assume that:*

1. *$\boldsymbol{\gamma}_0$ is in an open subset of the Euclidean space*
2. *$\boldsymbol{\gamma} \to \boldsymbol{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\gamma})$ is twice continuously differentiable $\forall(\boldsymbol{l}_i,\boldsymbol{a}_i)$*
3. *$E_{F_0}\left\|\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma}_0)\right\|_2^2 < \infty$*
4. *$E_{F_0}\left[\dot{\boldsymbol{\psi}}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma}_0)\right]$ exists and is non-singular*
5. *$\exists$ measurable integrable function $\ddot{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i)$ fixed such that $\ddot{\psi}_\gamma$ dominates the second partial derivatives of $\boldsymbol{\psi}_\gamma$ for all $\boldsymbol{\gamma}$ in a neighborhood of $\boldsymbol{\gamma}_0$.*

where $\boldsymbol{\gamma}_0$ are the true parameters of the propensity score model, and $\dot{\boldsymbol{\psi}}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\gamma})$ is the matrix of partial derivatives of $\boldsymbol{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. Then, $\sqrt{n}\left(\widehat{Y}_N^L(\alpha) - \boldsymbol{\mu}_0\right) \xrightarrow{d} N(0, W(\boldsymbol{\gamma}_0,\boldsymbol{\mu}_0))$, where

$$
\begin{aligned}
&W(\boldsymbol{\gamma}_0,\boldsymbol{\mu}_0) = V(\boldsymbol{\mu}_0) + A_{21}B_{11}^{-1}A_{21}^T + A_{21}B_{11}^{-1}B_{12} + \left(A_{21}B_{11}^{-1}B_{12}\right)^T,\\
&A_{21} = E\Big[\partial\psi_0/\partial\boldsymbol{\gamma}\ \ \partial\psi_1/\partial\boldsymbol{\gamma}\Big]^T,\ B_{11} = E\Big[\boldsymbol{\psi}_\gamma\boldsymbol{\psi}_\gamma^T\Big],\ B_{12} = E\Big[\boldsymbol{\psi}_\gamma\psi_0, \boldsymbol{\psi}_\gamma\psi_1\Big],
\end{aligned}
$$

evaluated at $(\boldsymbol{\gamma}_0,\boldsymbol{\mu}_0)$, $\psi_a = \psi_{a,\alpha}(\boldsymbol{Y}_i,\boldsymbol{A}_i,\boldsymbol{L}_i;\mu_0^a)$ and $V(\boldsymbol{\mu}_0)$ is that of Theorem 1.

$W(\boldsymbol{\gamma}_0,\boldsymbol{\mu}_0)$ can be easily estimated using $\widehat{W}\left(\widehat{\boldsymbol{\gamma}},\widehat{Y}_N^L(\alpha)\right)$, where $\widehat{W}(\boldsymbol{\gamma},\boldsymbol{\mu})$ is the matrix $W(\boldsymbol{\gamma},\boldsymbol{\mu})$ where all expectations are substituted with the empirical expectations. For example, $\widehat{B}_{11} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma})\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma})^T$.
Next, we derive the asymptotic distribution for $\widehat{\boldsymbol{\mu}}^{IE} = \left(\widehat{Y}_N^L(0;\alpha_0),\ \widehat{Y}_N^L(0;\alpha_1)\right)^T$ for the estimated propensity score from a correctly specified parametric model.

Theorem 3. *Denote* $\boldsymbol{\mu}_0^{IE} = \left(E_{F_0}\left[\overline{Y}_i^L(0,\alpha_0)\right],\ E_{F_0}\left[\overline{Y}_i^L(0,\alpha_1)\right]\right)^T$*, and assume that the Assumptions of Theorem 2 hold. Then,* $\sqrt{n}\left(\widehat{\boldsymbol{\mu}}^{IE} - \boldsymbol{\mu}_0^{IE}\right) \to N(0, Q(\boldsymbol{\gamma}_0, \boldsymbol{\mu}_0))$*, where*

$$
\begin{aligned}
&Q(\boldsymbol{\gamma}, \boldsymbol{\mu}) = D_{22} + C_{21}B_{11}^{-1}C_{21}^T + C_{21}B_{11}^{-1}D_{12} + \left(C_{21}B_{11}^{-1}D_{12}\right)^T \\
&D_{22} = Cov\left[\left(\overline{Y}_i^L(0,\alpha_1), \overline{Y}_i^L(0,\alpha_2)\right)^T\right],\ D_{12} = E\left[\boldsymbol{\psi}_\gamma \psi_{0,\alpha_1}, \boldsymbol{\psi}_\gamma \psi_{0,\alpha_2}\right], \\
&C_{21} = E\big[\partial\psi_{0,\alpha_1}/\partial\boldsymbol{\gamma}\ \ \partial\psi_{0,\alpha_2}/\partial\boldsymbol{\gamma}\big],
\end{aligned}
$$

and B_{11} *as in Theorem 2.*

3.4 Counterfactual distribution of cluster-average treatment propensity

In Section 3.2 we defined the individual average potential outcome for unit j in cluster i (and other estimands based on it) when the cluster-average propensity of treatment α is fixed to a counterfactual value. Those estimands correspond to quantities of interest in counterfactual worlds were one intervenes at the level of the cluster, but units within the cluster are still allowed to choose their own treatment. In this section, new individual average potential outcomes are defined, when the unit's treatment is set to a, but the cluster average propensity of treatment is not fixed to a specific value α but arises from a hypothesized distribution.

These estimands play an important role for policy interventions that occur at a high (vs. local) administrative level. For example, consider an observed distribution of cluster-average treatment propensity $\widehat{F}_\alpha$, and an intervention that takes place over all clusters incentivizing the increase of cluster treatment coverage. This intervention does not enforce a specific average propensity of treatment for each cluster separately, but leads to an overall shift in the distribution of cluster average propensity of treatment.

Let $F_\alpha(\cdot)$ denote the observed or a hypothesized distribution of cluster-average propensity of treatment. Then, define the F_α-individual potential outcome as

$$
\begin{aligned}
\overline{Y}_{ij}^L(a; F_\alpha) &= \int \overline{Y}_{ij}^L(a;\alpha)\ \mathrm{d}F_\alpha(\alpha) \qquad (3.6)\\
&= \sum_{\boldsymbol{s}\in\mathcal{A}(n_i-1)} Y_{ij}(A_{ij}=a, \boldsymbol{A}_{i,-j}=\boldsymbol{s}) \int P_{\alpha,L}(\boldsymbol{A}_{i,-j}=\boldsymbol{s}|A_{ij}=a, \boldsymbol{L}_i)\ \mathrm{d}F_\alpha(\alpha).
\end{aligned}
$$

Thus, $\overline{Y}_{ij}^{L}(a; F_\alpha)$ describes the average potential outcome of unit j in cluster i, for cluster average probability of treatment arising from F_α. Consequently, the F_α-group and population average potential outcomes are defined as

$$\overline{Y}_{i}^{L}(a; F_\alpha) = \frac{1}{n_i} \sum_{j=1}^{n_i} \overline{Y}_{ij}^{L}(a; F_\alpha)$$

and

$$\overline{Y}^{L}(a; F_\alpha) = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_{i}^{L}(a; F_\alpha)$$

accordingly. Although the above estimands are well-defined for a distribution F_α different than the observed one, F_α needs to have overlapping support with the empirical distribution $\widehat{F}_\alpha$ in order to reliably estimate such quantities.

Similar arguments to the ones in Section 3.3 lead to estimators of the F_α-group and population average potential outcome as

$$\widehat{Y}_{i}^{L}(a; F_\alpha) = \int \widehat{Y}_{i}^{L}(a; \alpha) \mathrm{d}F_\alpha = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{F_\alpha, L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij}, \tag{3.7}$$

$$\widehat{Y}^{L}(a, F_\alpha) = \int \widehat{Y}^{L}(a; \alpha) \mathrm{d}F_\alpha = \frac{1}{N} \sum_{i=1}^{n_i} \widehat{Y}_{i}^{L}(a; F_\alpha) \tag{3.8}$$

accordingly, where $P_{F_\alpha, L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i) = \int P_{\alpha, L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i) \, \mathrm{d}F_\alpha(\alpha)$.

Even though direct effect estimands based on the F_α-population average potential outcome can easily be defined as $DE(F_\alpha) = \overline{Y}^{L}(1; F_\alpha) - \overline{Y}^{L}(0; F_\alpha)$, the contrast of F_α-population average potential outcomes is more interesting for the indirect effect. For two hypothesized distributions of cluster-average propensity of treatment F_α^1, F_α^2, define

$$IE\left(F_\alpha^1, F_\alpha^2\right) = \overline{Y}^{L}\left(0; F_\alpha^2\right) - \overline{Y}^{L}\left(0; F_\alpha^1\right).$$

Then, $IE\left(F_\alpha^1, F_\alpha^2\right)$ represents the expected outcome change for control units when the distribution of cluster-average propensity of treatment changes from F_α^1 to F_α^2.

Assume that F_α^1, F_α^2 represent discrete distributions with values $\alpha_1, \alpha_2, \ldots, \alpha_K \in (0, 1)$ and probability p_{1k} and p_{2k} of assigning value α_k to a cluster accordingly, such that $\sum_{k=1}^{K} p_{jk} = 1, \ j = 1, 2$. Then,

$$\overline{Y}\left(0, F_\alpha^j\right) = \sum_{k=1}^{K} p_{jk} \overline{Y}(0, \alpha_k) \ \Rightarrow \ IE\left(F_\alpha^1, F_\alpha^2\right) = \sum_{k=1}^{K} (p_{2k} - p_{1k}) \overline{Y}(0, \alpha_k).$$

Clearly, a consistent estimator for $IE\left(F_{\alpha}^{1}, F_{\alpha}^{2}\right)$ is

$$\widehat{IE}\left(F_{\alpha}^{1}, F_{\alpha}^{2}\right) = \sum_{k=1}^{K}(p_{2k} - p_{1k})\widehat{Y}(0, \alpha_k).$$

Acquiring the asymptotic distribution of $\widehat{IE}\left(F_{\alpha}^{1}, F_{\alpha}^{2}\right)$ is straightforward following similar arguments to the ones in Theorem 3 to acquire the asymptotic distribution of $\big(\widehat{Y}(0, \alpha_1), \widehat{Y}(0, \alpha_2), \ldots, \widehat{Y}(0, \alpha_K)\big)^T$ and applying the multivariate delta method.

3.5 Simulations

We generate a fixed population of 2,000 clusters including 14 to 18 units each, resulting to a total of 31,553 units. Four independent $N(0, 1)$ covariates were generated, and are denoted as L_1, L_2, L_3, L_4. For every individual in the population (unit j in cluster i), the potential outcomes under all possible treatment allocations were generated, following a model $Y \sim$ Bernoulli$(\text{expit}(l_Y))$ where

$$l_Y = 0.5 - 0.6a - 1.4\frac{a+k}{n_i} - 0.098L_{1ij} - 0.145L_{2ij} + 0.1L_{3ij} + 0.3L_{4ij} + 0.351a\frac{a+k}{n_i}, \quad (3.9)$$

$L_{1ij}, L_{2ij}, L_{3ij}, L_{4ij}$ are the values of the covariates for observation j of cluster i, a is the individual treatment, k is the number of treated neighbors, and $(a+k)/n_i$ is the percentage of units in the cluster that are treated.

3.5.1 A simulated data set

The simulations test the operating characteristics of the estimator in (3.5) using the true and estimated propensity score in terms of the re-sampling of the observed treatment vector. Specifically, each simulated dataset includes the whole population, but a different set of potential outcomes is observed according to a treatment vector generated as $A_{ij} \sim$ Bernoulli$(\text{expit}(l_A))$ where

$$l_A = -0.2 + b_i + 0.3L_{1ij} - 0.15L_{2ij} + 0.2L_{3ij} - 0.18L_{4ij},\ b_i \sim N(0, 0.5^2). \quad (3.10)$$

Once the observed treatment is generated, the observed outcome is the corresponding value of the potential outcomes. Clusters with all treated or all control units are dropped.

$\widehat{Y}_i(a;\alpha)$ is estimated from (3.4) for counterfactual treatment allocation described in Section 3.5.2, and

$$f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i;\boldsymbol{\gamma}) = \int \prod_{j=1}^{n_i} f_e\left(A_{ij}|L_{ij},\delta_0,\beta_i,\boldsymbol{\delta}\right)\phi\left(\beta_i|\sigma_\beta^2\right)\mathrm{d}\beta_i,$$

where

$$f_e\left(A_{ij}|L_{ij},\delta_0,\beta_i,\boldsymbol{\delta}\right) = \mathrm{expit}\left(\delta_0+b_i+L_{ij}^T\boldsymbol{\delta}\right)^{A_i}\left[1-\mathrm{expit}\left(\delta_0+b_i+L_{ij}^T\boldsymbol{\delta}\right)\right]^{1-A_i},$$

$L_{ij}^T = (L_{1ij}, L_{2ij}, L_{3ij}, L_{4ij})$, $\phi\left(\cdot;\sigma_\beta^2\right)$ the density of a $N\left(0,\sigma_\beta^2\right)$, and $\boldsymbol{\gamma} = \left(\delta_0,\boldsymbol{\delta},\sigma_\beta^2\right)$ known and equal to the coefficients in (3.10), or the maximum likelihood estimates from the correctly specified propensity score model.
We calculate the population average potential outcomes, direct and indirect effects, and the corresponding asymptotic variances.

3.5.2 Covariate-dependent counterfactual treatment allocation

The counterfactual treatment allocation $P_{\alpha,L}$ is allowed to depend on the same covariates that are included in the observed propensity score, using the log odds coefficients used to generate the observed treatment. Specifically, for a fixed $\alpha \in (0,1)$,

$$\mathrm{logit}P_{\alpha,L}(A_{ij}=1|L_{ij}) = \xi_i^\alpha + 0.3L_{1ij} - 0.15L_{2ij} + 0.2L_{3ij} - 0.18L_{4ij},$$

for ξ_i^α satisfying $\frac{1}{n_i}\sum_{j=1}^{n_i} P_{\alpha,L}(A_{ij}=1|L_{ij}) = \alpha$. (Description of how ξ_i^α is calculated can be found in Section 3.8.6.)

3.5.3 Calculating the true average potential outcomes

For every observation j in cluster i, the individual average potential outcome for individual treatment a and for cluster-average propensity of treatment α is calculated based on (3.1). Based on the individual average potential outcome, the true group and population average potential outcome are calculated according to (3.2), (3.3).

3.5.4 Simulation results

We present results for values of alpha between 0.3 and 0.6 corresponding to the 20^{th} and 80^{th} quantiles of the distribution of the observed treatment proportions across clusters

and simulated data sets. As expected, the estimator based on the true propensity score is unbiased, while the estimator based on the estimated propensity score, which is consistent but not unbiased, indicates small biases. Figure 3.1 shows the mean estimate across 500 simulated data sets for the population average potential outcome for $a = 1$ (results were similar for $a = 0$), direct and indirect effect, and Table 3.1 shows the coverage range over different values of α of the population average potential outcome, direct and indirect effect.

Moreover, Figure 3.5 in Section 3.8.1 shows the mean of the estimated variance based on the asymptotic results, against the variance of the estimates calculated over the 500 simulated data sets. The points (one for each value of α) lie mostly on the 45 degree

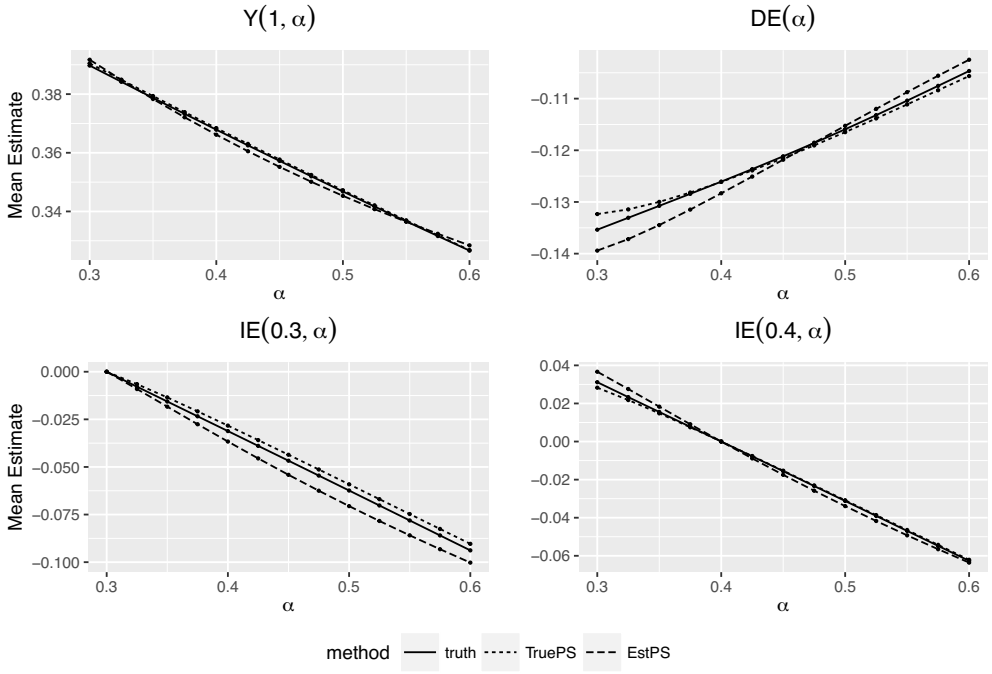


Figure 3.1: Mean estimate of population average potential outcome, direct, and indirect effect over 500 simulated data sets for the true or the correctly specified propensity score.

Table 3.1: Range of percent coverage over 500 simulated data sets for the population average potential outcome, direct and indirect effects for the true or the correctly specified propensity score.

	$\overline{Y}^L(0;\alpha)$	$\overline{Y}^L(1;\alpha)$	$DE^L(\alpha)$	$IE^L(\alpha_1,\alpha_2)$
True PS	94.4 - 96.2 %	94.4 - 97.2 %	94.4 - 95.6 %	92 - 97.2 %
Estimated PS	86.6 - 96 %	92.8 - 96.6 %	91.6 - 96 %	78.4 - 95.6 %

line indicating that, on average, the variance based on the asymptotic theory is a good approximation of the true variance.

3.6 Application: Effectiveness of Power Plant Emissions Controls for Reducing Ambient Ozone Pollution

Limited literature exists in the evaluation of U.S. air pollution regulations in a causal inference framework. Power plant regulations for the reduction of NO_x emissions have been predicated on the knowledge that reducing NO_x emissions would lead to a subsequent reduction in ambient ozone. Among various NO_x emission reduction strategies, SCR and SNCR are believed to be the most effective in reducing emissions. While work in Papadogeorgou et al. (2018) corroborated this effectiveness of SCR and SNCR in an analysis for NO_x emissions, the analysis of ambient ozone pollution in that paper ignores the possibility of interference and estimates a null effect on ambient ozone. However, interference is a key component in the study of air pollution: ambient pollution concentrations near a power plant will depend on the treatment levels of other nearby power plants. Causal estimands tailored to settings of interference can answer important questions related to the effectiveness of interventions in the presence of long-range pollution transport.

We use the same data as in Papadogeorgou et al. (2018) to estimate direct and indirect effects of SCR/SNCR against alternatives on ambient ozone under realistic counterfactual programs. The publicly-available data set includes 473 coal or gas burning power generating facilities in the U.S. operating during June, July and August 2004, with covariate information on power plant characteristics, weather and demographic information of the surrounding areas. For every power plant, the value of ozone is calculated as the average across EPA monitoring locations within 100km of the 4^{th} highest ozone measurements. See Papadogeorgou et al. (2018) for a full description of the data set and linkage.

Power plant facilities are grouped into 50 clusters according to Ward's agglomerative clustering method (Ward, 1963) based on coordinates. The grouping and treatment of facilities are depicted in Figure 3.2. 10 out of 50 clusters were excluded from the analysis because they included only control power plants.

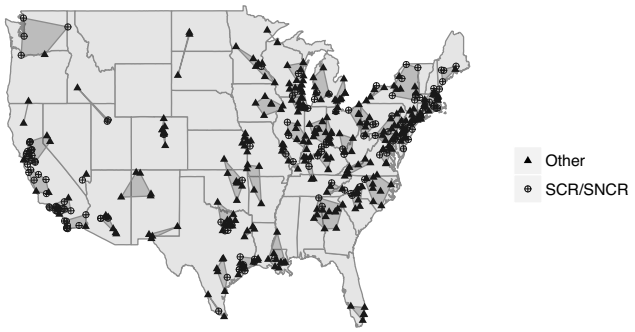


Figure 3.2: Treated (SCR/SNCR) and control (Other) power plant facilities during June, July, August of 2004. Shaded areas depict the interference clusters according to the agglomerative clustering method.

3.6.1 Plausibility of the ignorability and positivity assumption

While regulatory programs provide incentives to install emission-control technologies, power plants have latitude to select which (if any) technology to adopt. Such decisions are largely determined by the plant's characteristics such as plant size and operating capacity, as well as by factors related to local or regional air pollution incentives that are influenced by area-level characteristics such as population density and urbanicity. To capture such factors, 18 covariates are included in the data set describing power plant, weather, and demographic characteristics, based on which ignorability is expected to hold. The variability in the observed proportion of treated power plants across clusters provides an additional indication that the positivity assumption is plausible. Based on these covariates, the propensity score was modeled as in Papadogeorgou et al. (2018) augmented with a cluster-specific random effect

$$\text{logit}P(A_{ij} = 1|L_{ij}) = \delta_0 + b_i + L_{ij}^T\boldsymbol{\delta},\ b_i \sim N(0, \sigma_b^2). \tag{3.11}$$

3.6.2 Counterfactual treatment allocation for the installation of SCR/SNCR emission control technologies

Recall from Section 3.2.3 that $P_{\alpha,L}$ governing treatment assignment in the counterfactual allocation programs of interest must be specified. To specify counterfactual treatment allocations that reflect realistic relationships between covariates and the propensity to

adopt treatment, we specify $P_{\alpha,L}$ such that the log-odds of treatment installation related to individual covariates are as observed in the propensity score model for the observed treatment in (3.11). Even though this choice of $P_{\alpha,L}$ depends on the data through the estimated log-odds, the corresponding estimands are well-defined and the asymptotic results are valid for $P_{\alpha,L}$ fixed across replications of the sampling or an increasing number of clusters.

Values of α were considered between the 20^{th} and 80^{th} quantiles of the observed cluster treatment proportions, corresponding to $\alpha \in [0.141, 0.508]$. Figure 3.3 shows the population direct effect $DE(\alpha)$, and population indirect effect $IE(\alpha_1, \alpha_2)$ for a subset of values of α_1 (for presentation simplicity). The direct effect is significantly negative for all values of α, but has a somewhat increasing trend, implying that in a world where the average probability of SCR/SNCR among power plants in a cluster is fixed, the installation of SCR/SNCR at one power plant would lead to significant reductions in ozone concentrations in the surrounding area, but these reductions are smaller when the cluster average propensity of treatment is high (larger number of treated neighbors).

The indirect effect is, in a way, a measure of pollution transport since it quantifies the effect of changes in the cluster average propensity of treatment on ozone concentrations near control power plants. For all values of α_1, $IE(\alpha_1, \alpha_2)$ is decreasing in α_2, and almost all contrasts considered were significant at the 0.05 significance level. The decreasing trend in $IE(\alpha_1, \alpha_2)$ for a fixed value of α_1 implies that higher cluster-average SCR/SNCR

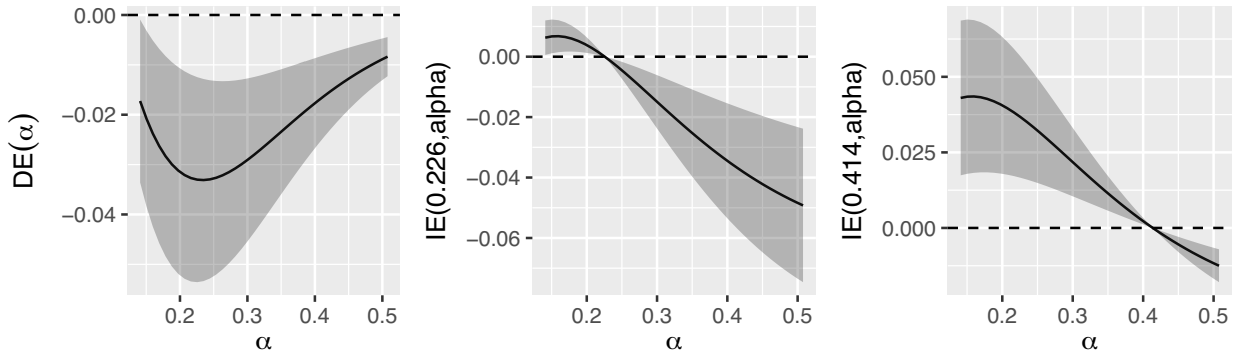


Figure 3.3: Direct effect of control versus treated power plants on ozone concentrations as a function of α, and indirect effect where the first value of α is fixed to a specific value. Ozone is measured in parts per million.

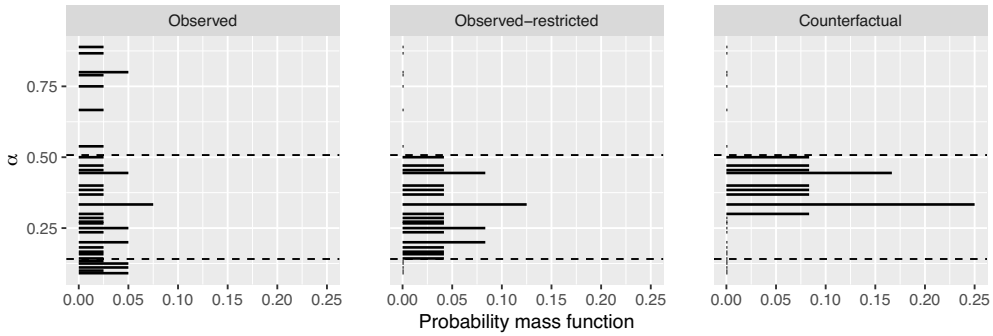


Figure 3.4: Observed cluster treatment proportions ("Observed"), and two discrete hypothesized distributions of cluster-average probability of treatment. One corresponds to the observed restricted within the 20^{th} and 80^{th} quantiles of the observed cluster treatment proportions ("Observed-restricted"), and the other one ("Counterfactual") corresponds to the observed (or the Observed-restricted) further restricted between the 50^{th} and 80^{th} quantiles of the observed cluster treatment proportions.

propensity leads to further decrease in ambient ozone concentrations in the surrounding area of control power plants.

Next, we considered estimating the effect of hypothesized federal regulations that would shift the distribution of cluster-average propensity of treatment. F_α^1 (F_α^2) was assumed to be a discrete distribution within the 20^{th} (50^{th}) and 80^{th} quantiles of the observed cluster-treatment proportions. In Figure 3.4, we show the empirical probability mass function, as well as the two counterfactual treatment allocations. $IE\left(F_\alpha^1, F_\alpha^2\right)$ was estimated to be -0.0162 parts per million (95% CI: -0.0252 to -0.007) implying that federal regulations that encourage the installation of SCR/SNCR and would lead to cluster average treatment probabilities like F_α^2 would reduce ambient ozone concentrations in the surrounding areas of control power plants by 0.0162 parts per million compared to similar regulations that would lead to F_α^1. For reference, these effect estimates can be compared against the national ozone air quality standard of 0.07 parts per million.

We explored the sensitivity of the data application results to the choice of hierarchical clustering method and number of clusters, and saw that the qualitative results for the effectiveness of SCR/SNCR emission reduction technologies are mostly consistent. We further estimated the estimands of Tchetgen Tchetgen and VanderWeele (2012) that assume manipulation of individual power-plant treatment propensities and ignore the fact

that covariates can be important predictors of the treatment received by power plants. These estimators returned results with similar pattern as the ones in Figure 3.3, but of smaller magnitude. Therefore, the potential benefits of SCR/SNCR would be understated, if counterfactual scenarios of independent and identically distributed treatment assignment was considered. These results can be found in Section 3.8.2, along with links to the publicly available data set, R package and scripts.

3.7 Discussion

Analyzing data in the context of interference disentangles the effect of the individual treatment from the treatment of one's neighbors. New estimands in the presence of interference were proposed for counterfactual strategies that manipulate treatment at the cluster-level, or at the level of population of clusters. These new estimands represent scenarios where individual treatment in the counterfactual world is allowed to depend on covariates and the treatment of one's neighbors. Such estimands are relevant for public health interventions that do not manipulate treatment at the unit level.

For the estimands referring to interventions at the population level, the counterfactual distribution F_α represented the distribution of the cluster-average propensity of treatment, and each cluster was assumed to be equally likely to receive α from F_α. However, F_α could be alternatively defined to depend on cluster-level covariates that act as predictors of cluster-average propensity of treatment.

Consistent estimators were proposed for which the asymptotic distribution was derived. These estimators were employed in the comparative effectiveness of power plant emission control strategies on ambient ozone, and showed the potential of a set of emission reduction technologies in reducing ozone concentrations. These results are more in line with subject-matter knowledge than results from a previous study that assumed no interference. Through comparison with existing estimands in the literature, this analysis also highlighted the potential gains of the proposed estimands in a setting where relevant counterfactual treatment allocations depend on covariates, showing that existing estimands would understate the potential benefits of treatment.

Even though the data application showed the potential for causal inference methods for interference to lead to important results in air pollution research, there are several limitations to the analysis. First of all, the number of clusters was low, raising questions for the appropriateness of use of asymptotic distributions to acquire variance estimates. Furthermore, in air pollution studies the assumption of partial interference may be violated, since pollution from one power plant can travel long distances and affect ozone concentrations within a different cluster. However, we considered it important to analyze air quality data, which in our knowledge have not been previously analyzed to acknowledge interference. Despite the approximation entailed by the partial interference assumption, we believe this is an important advance for studies of air pollution interventions, and methods relaxing the assumption of partial interference for unknown networks are the topic of future research.

3.8 Appendix

3.8.1 Simulation results

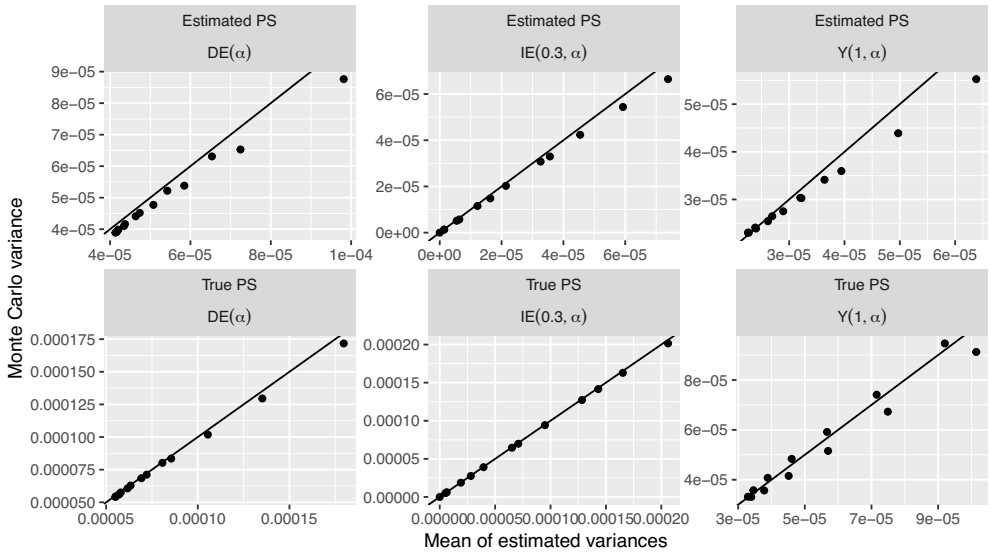


Figure 3.5: Mean estimated variance from the asymptotic distribution, and Monte Carlo variance of the estimates. The diagonal lines correspond to the 45 degree line, and each point corresponds to a value of α.

3.8.2 Data application

Link to the publicly available data, the R package implementing the estimators, and scripts replicating the results of the data analysis are available at `https://osf.io/7dp8c/`.

Sensitivity of data application results to the choice of clustering

Rows correspond to the direct effect $DE(\alpha)$ and the indirect effects $IE(\alpha_1, \alpha_2)$ for $\alpha_1 \in \{0.32, 0.41\}$. Columns correspond to the clustering method and correspond to Ward's Ward (1963) method for 30 and 70 clusters, and complete clustering with 50 clusters. The increasing trend in the indirect effect persists for all clustering specifications, while the direct effect results are sensitive to the specification of the hierarchical clustering method used.

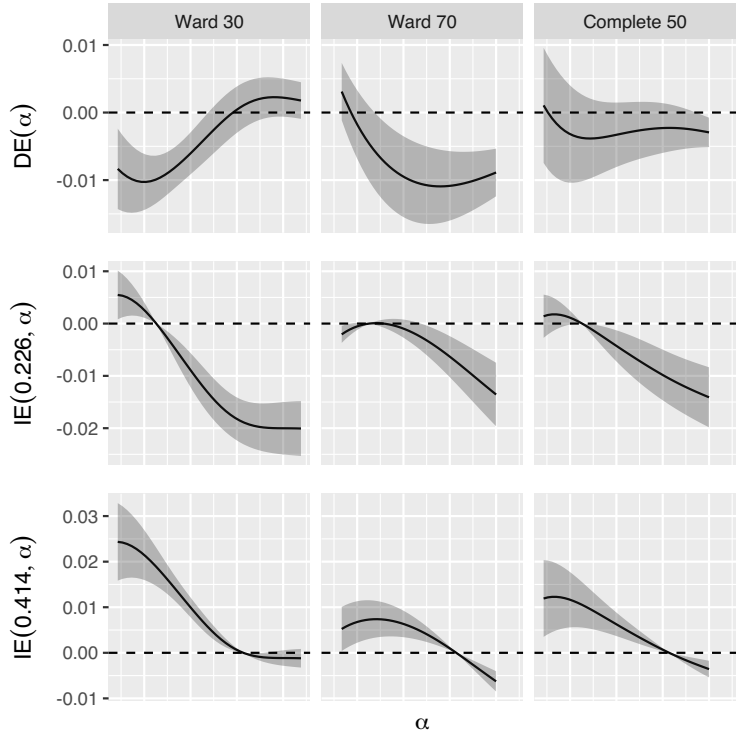


Figure 3.6: Direct and indirect effect of SCR/SNCR on ambient ozone using different clustering of power plants. Methods for clustering from left to right include Ward's method for 30 and 70 clusters, and complete clustering using 50 clusters.

Estimates and confidence intervals of the direct and indirect effect under an independent Bernoulli treatment allocation

We estimate the direct and indirect effects under the counterfactual scenario found in Tchetgen Tchetgen and VanderWeele (2012) that considers assigning treatment to units independently and with equal probability. Confidence intervals are based on asymptotic distributions derived as in Theorems 2, 3.

The results' pattern is similar to the ones in Figure 3.3, but all estimates are closer to 0. This implies that basing our inferences on estimands in unrealistic counterfactual treatment allocation programs would underestimate the potential benefits of SCR/SNCR.

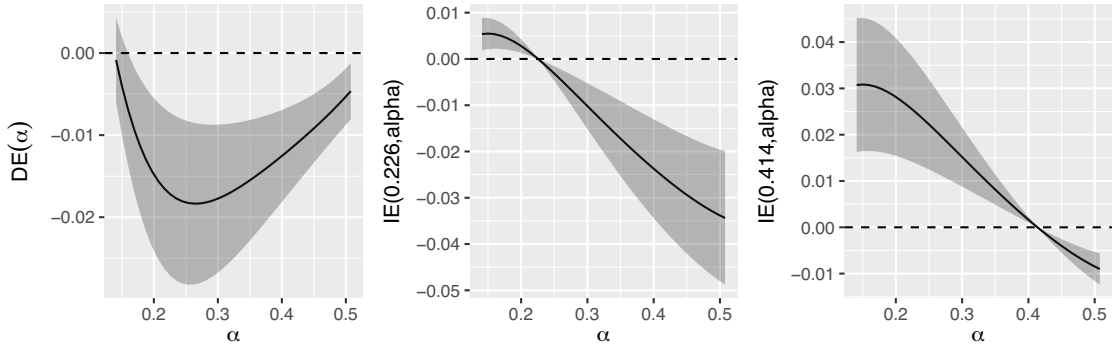


Figure 3.7: Direct and indirect effect estimates and confidence intervals for the estimands defined in Tchetgen Tchetgen and VanderWeele (2012).

3.8.3 Proofs of unbiasedness, consistency and asymptotic normality

Unbiasedness

Theorem 4. *If $f_{\boldsymbol{A}|\boldsymbol{L},i}(\cdot|\boldsymbol{L}_i)$ is known, and Assumptions 1, 2 hold, then $\widehat{Y}_i^L(a;\alpha)$ is an unbiased estimator for the group average potential outcome, and $\widehat{Y}^L(a;\alpha)$ is an unbiased estimator of the population average potential outcome for individual treatment a and cluster average propensity of treatment α.*

Proof. All expectations are taken with respect to the conditional distribution $\boldsymbol{A}_i|\boldsymbol{L}_i, \boldsymbol{Y}_i(\cdot)$, where $\boldsymbol{Y}_i(\cdot)$ are all the potential outcomes for all units in cluster i. Y_ij, $\boldsymbol{Y}_i$ denote the observed individual outcome, and the vector of observed outcomes in cluster i accordingly.

$$E[\widehat{Y}_i^L(a;\alpha)] = \frac{1}{n_i}\sum_{j=1}^{n_i} E\left(\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij}\right)$$

$$
\begin{aligned}
=&\frac{1}{n_i}\sum_{j=1}^{n_i} E\left(\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij}(\boldsymbol{A}_i)\right)\\
=&\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\boldsymbol{s}\in\mathcal{A}(n_i)} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}=\boldsymbol{s}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i=\boldsymbol{s}|\boldsymbol{L}_i)} I(s_{ij}=a)Y_{ij}(\boldsymbol{s})P(\boldsymbol{A}_i=\boldsymbol{s}|\boldsymbol{L}_i,\boldsymbol{Y}_i(\cdot))\\
=&\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\boldsymbol{s}\in\mathcal{A}(n_i)} f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}=\boldsymbol{s}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)I(s_{ij}=a)Y_{ij}(\boldsymbol{s})\\
&\text{(From Assumption 2 } P(\boldsymbol{A}_i=\boldsymbol{s}|\boldsymbol{L}_i,\boldsymbol{Y}_i(\cdot))=P(\boldsymbol{A}_i=\boldsymbol{s}|\boldsymbol{L}_i)=f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i=\boldsymbol{s}|\boldsymbol{L}_i)\text{.)}\\
=&\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\boldsymbol{s}\in\mathcal{A}(n_i-1)} f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}=\boldsymbol{s}|A_{ij}=a,\boldsymbol{L}_i,\alpha)Y_{ij}(a_{ij}=a,a_{i,-j}=\boldsymbol{s})=\overline{Y}_i^L(a;\alpha).
\end{aligned}
$$

By linearity of expectations, the proof for the population average potential outcome is trivial. □

Proofs of asymptotic results for known propensity score

For notational simplicity, denote $\tilde{\boldsymbol{O}}_i=(\boldsymbol{A}_i,\boldsymbol{L}_i)$, $\boldsymbol{O}_i=(\boldsymbol{Y}_i,\boldsymbol{A}_i,\boldsymbol{L}_i)$, $\tilde{\boldsymbol{o}}_i=(\boldsymbol{a}_i,\boldsymbol{l}_i)$, and $\boldsymbol{o}_i=(\boldsymbol{y}_i,\boldsymbol{a}_i,\boldsymbol{l}_i)$. Also, denote as F_0 the distribution of $(\boldsymbol{Y}_i(\cdot),\boldsymbol{A}_i,\boldsymbol{L}_i)$ in the superpopulation. Consider the estimating equation $\Psi_N(\mu)=\sum_{i=1}^N\psi_{a,\alpha}(\boldsymbol{O}_i;\mu)=0$, where

$$
\psi_{a,\alpha}(\boldsymbol{O}_i;\mu)=\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right)-\mu.
$$

It is easy to see that the solution to this equation is $\widehat{\mu}=\widehat{Y}_N^L(a;\alpha)$:

$$
\begin{aligned}
&\sum_{i=1}^N\left[\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right)-\mu\right]=0 \iff\\
&\sum_{i=1}^N\widehat{Y}_i^L(a;\alpha)=N\mu \iff \widehat{\mu}=\frac{1}{N}\sum_{i=1}^N\widehat{Y}_i^L(a;\alpha)=\widehat{Y}_N^L(a;\alpha)
\end{aligned}
$$

If μ_0^a is the solution to $\Psi_0(\mu)=\int\psi_{a,\alpha}(\boldsymbol{O}_i;\mu)\mathrm{d}F_0(\boldsymbol{o}_i)=0$. Then,

$$
\begin{aligned}
&\int\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}-\mu_0^a\right]\mathrm{d}F_0(\boldsymbol{o}_i)=0 \iff\\
&\mu_0^a=E_{F_0}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right]=E_{F_0}\left[\overline{Y}_i^L(a;\alpha)\right]
\end{aligned}
$$

Proof of Theorem 1. First, we will show that $\widehat{Y}_N^L(\alpha)$ is consistent for $\boldsymbol{\mu}_0$. For this proof, we use an alteration of Lemma A in section 7.2.1 of Serfling (1980).

Note that $\psi_{a,\alpha}(\boldsymbol{O}_i;\mu)$ is monotone in μ with $\dot{\psi}_{a,\alpha}(\boldsymbol{O}_i;\mu) = \frac{\partial}{\partial\mu}\psi_{a,\alpha}(\boldsymbol{O}_i;\mu) = -1 < 0$. Therefore, $\Psi_N(\mu), \Psi_0(\mu)$ are also monotone in μ (implying uniqueness of their roots). From the strong law of large numbers we have that $\Psi_N(\mu) \xrightarrow{a.s.} \Psi_0(\mu)$. From this, we have that:

$$|\Psi_0(\widehat{\mu}) - \Psi_0(\mu_0)| = |\Psi_0(\widehat{\mu}) - \Psi_n(\widehat{\mu})| \leq \sup_{\mu} |\Psi_0(\mu) - \Psi_n(\mu)| \to 0,$$

which, by the uniqueness of the roots for Ψ_0, Ψ_N, implies $\widehat{\mu} \xrightarrow{a.s.} \mu_0$, and $\widehat{Y}_N^L(a;\alpha) \xrightarrow{a.s.} E_{F_0}\left[\overline{Y}_i^L(a;\alpha)\right]$.

From basic probability laws we have that $(X_n, Y_n)^T \xrightarrow{a.s.} (X, Y)^T$ if and only if $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$. We showed that the individual components converge almost surely to their limit, and therefore

$$\left(\widehat{Y}_N^L(0,\alpha), \widehat{Y}_N^L(1;\alpha)\right)^T \xrightarrow{a.s} \left(E_{F_0}\left[\bar{Y}_i^L(0,\alpha)\right], E_{F_0}\left[\bar{Y}_i^L(1,\alpha)\right]\right)^T,$$

which also establishes convergence in probability.

Now we will show that $\widehat{Y}_N^L(a;\alpha)$ has an asymptotically univariate normal distribution, for $a = 0, 1$, and afterwards we extend this to showing that $\widehat{Y}_N^L(\alpha)$ has an asymptotically bivariate normal distribution.

Univariate result

Based on the above, $\widehat{Y}^L(a,\alpha) \xrightarrow{p} E_{F_0}\left[\overline{Y}_i^L(a,\alpha)\right] = \mu_0^a$. Theorem A in section 7.2.2 of Serfling (1980) requires:

(i) μ_0^a is an isolated root of $\Psi_0(\mu) = 0$ and $\psi_{a,\alpha}(\cdot;\mu)$ is monotone in μ. (Shown above)

(ii) $\Psi_0(\mu)$ is differentiable at μ_0^a with $\Psi_0'(\mu_0^a) \neq 0$.

(iii) $\int \psi_{a,\alpha}^2(\boldsymbol{o}_i;\mu)\mathrm{d}F_0(\boldsymbol{o}_i)$ is finite in a neighborhood of μ_0^a.

Proof of (ii)

$$\begin{aligned}\Psi_0(\mu) =& \int \left[\frac{1}{n_i}\sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} - \mu\right]\mathrm{d}F_0(\boldsymbol{o}_i) \\ =& \frac{1}{n_i}\int \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij}\,\mathrm{d}F_0(\boldsymbol{o}_i) - \mu\end{aligned}$$

So Ψ_0 is linear in μ and therefore differentiable everywhere, with $\Psi_0'(\mu) = -1 \neq 0$.

Proof of (iii)

Consider a neighborhood of μ_0^a of the form $(\mu_0^a - \epsilon, \mu^a + \epsilon)$, for some $\epsilon > 0$. Then,

$$\begin{aligned}
&\int \psi_{a,\alpha}^2(\boldsymbol{O}_i;\mu)\mathrm{d}F_0(\boldsymbol{o}_i) \\
&= \int \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} - \mu \right)^2 \mathrm{d}F_0(\boldsymbol{o}_i) \\
&= \int \left| \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} - \mu \right|^2 \mathrm{d}F_0(\boldsymbol{o}_i)
\end{aligned}$$

I will show that $\left| \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} - \mu \right|$ is bounded by a constant c in a neighborhood of μ_0 and therefore the integral is bounded by c^2.

$$\begin{aligned}
&\left| \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} I(A_{ij}=a)Y_{ij} - \mu \right| \\
&\leq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} |Y_{ij}| + |\mu| \\
&\leq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|Y_{ij}|}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} + |\mu| \\
\Rightarrow &\int \psi_{a,\alpha}^2(\boldsymbol{O}_i;\mu)\mathrm{d}F_0(\boldsymbol{o}_i) \leq |\mu| + (n_i)^{-1} \sum_{j=1}^{n_i} E_{F_0}\left[\frac{|Y_{ij}|}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)} \right] \\
&\leq |\mu| + (n_i)^{-1} \sum_{j=1}^{n_i} \sqrt{E_{F_0}(Y_{ij}^2) E_{F_0}(f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)^{-2})} < |\mu| + M\rho^{-1} = c \qquad (3.12)
\end{aligned}$$

We have shown that the conditions of Theorem A (section 7.2.2 of Serfling (1980)) are satisfied, and therefore

$$\sqrt{n}\left(\widehat{Y}_N^L(a;\alpha) - E_{F_0}\left[\overline{Y}_i^L(a,\alpha) \right] \right) \xrightarrow{d} N(0,\sigma^2),$$

where $\sigma^2 = E\left[\psi_{a,\alpha}^2\left(\boldsymbol{O}_i;\mu_0^a\right) \right]$, since $\Psi_0'(\mu_0) = -1$.

Bivariate result

We will use Theorem 5.41 of van der Vaart (1998). The assumptions of this theorem are the so-called "classical" conditions, and are stricter than necessary to prove asymptotic

normality. However, this theorem is often used in practice, since the conditions are sometimes easy to prove, as they are here.

We denote $\psi(\boldsymbol{o}_i;\boldsymbol{\mu}) = (\psi_{0,\alpha}(\boldsymbol{o}_i;\mu^0),\psi_{1,\alpha}(\boldsymbol{o}_i;\mu^1))^T$, for $\boldsymbol{\mu} = (\mu^0,\mu^1)$,. and $\Psi_n(\boldsymbol{\mu})$, $\Psi_0(\boldsymbol{\mu})$ similarly as above, but for the vector ψ.

It was shown that $\boldsymbol{\mu}_0$ satisfies $\Psi_0(\boldsymbol{\mu}) = 0$, and that $\widehat{Y}_N^L(\alpha)$ is a consistent estimator of $\boldsymbol{\mu}_0$. In order to apply Theorem 5.41, we show that

(i) the function $\boldsymbol{\mu} \to \psi(\boldsymbol{o}_i;\boldsymbol{\mu})$ is twice continuously differentiable for every vector $\boldsymbol{o}_i$,

(ii) $E_{F_0}\|\psi(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\|_2^2 < \infty$ (where $\|\cdot\|_2$ is the 2-norm $\|(v_1,v_2,\ldots,v_n)\|_2 = (v_1^2+v_2^2+\cdots+v_n^2)^{1/2}$,

(iii) The matrix $E_{F_0}\left[\dot{\psi}(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\right]$ exists and is nonsingular, and

(iv) $\exists$ fixed integrable function $\ddot{\psi}(\boldsymbol{o}_i)$ such that $\ddot{\psi}$ dominates the second order partial derivatives of ψ $\forall\boldsymbol{\mu}$ in a neighborhood of $\boldsymbol{\mu}_0$.

Proof of (i). It has already been shown that $\psi_{a,\alpha}(\boldsymbol{o}_i;\mu)$ is linear in μ and therefore twice continuously differentiable with respect to μ for every vector $(\boldsymbol{o}_i)$.

Proof of (ii).

$$\begin{aligned}
E_{F_0}\|\psi(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\|_2^2 =& E\left\{\sum_{a\in\{0,1\}}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}-\mu_0^a\right]^2\right\}\\
=& \sum_{a\in\{0,1\}}E\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i,\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}-\mu_0^a\right]^2\\
\leq& 2c^2 \qquad \text{(because of (3.12))}
\end{aligned}$$

Proof of (iii).

$$\dot{\psi}(\boldsymbol{o}_i;\boldsymbol{\mu}) = \begin{pmatrix}\frac{\partial\psi_{0,\alpha}(\boldsymbol{o}_i;\mu^0)}{\partial\mu^0} & \frac{\partial\psi_{0,\alpha}(\boldsymbol{o}_i;\mu^0)}{\partial\mu^1}\\ \frac{\partial\psi_{1,\alpha}(\boldsymbol{o}_i;\mu^1)}{\partial\mu^0} & \frac{\partial\psi_{1,\alpha}(\boldsymbol{o}_i;\mu^1)}{\partial\mu^1}\end{pmatrix} = \begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix} = -I_2 < \infty \text{ and non-singular}, \quad (3.13)$$

where the diagonal elements of partial derivatives are calculated in the proof of consistency, and the non-diagonal elements are clearly 0 since the functions do not include the corresponding components of $\boldsymbol{\mu}$.

Proof of (iv). Based on equation (3.13), we have that all second order derivatives are equal to 0, and are therefore dominated by the integrable function $\ddot{\psi}(\boldsymbol{o}_i) = 0$.
From Theorem 5.41 of van der Vaart (1998), we have that

$$\sqrt{n}\left(\widehat{Y}_N^L(\alpha) - \boldsymbol{\mu}_0\right) = -\left(E\left[\dot{\psi}(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\right]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{N}\psi(\boldsymbol{O}_i;\boldsymbol{\mu}_0) + o_P(1)$$
$$\Rightarrow \sqrt{n}\left(\widehat{Y}_N^L(\alpha) - \boldsymbol{\mu}_0\right) \xrightarrow{d} N\left(0, A(\boldsymbol{\mu}_0)^{-1}V(\boldsymbol{\mu}_0)\left[A(\boldsymbol{\mu}_0)^{-1}\right]^T\right),$$

where $A(\boldsymbol{\mu}_0) = E\left[-\dot{\psi}(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\right] = I_2$ (from 3.13), and $V(\boldsymbol{\mu}_0) = E\left[\psi(\boldsymbol{O}_i;\boldsymbol{\mu}_0)\psi(\boldsymbol{O}_i;\boldsymbol{\mu}_0)^T\right]$. □

Example 1. We provide an example of the application of the delta method on the result of Theorem 1. Consider the direct effect defined as $\overline{DE}^L(\alpha) = \overline{Y}^L(0;\alpha) - \overline{Y}^L(1;\alpha)$. Then $\widehat{\overline{DE}}^L(\alpha) = \widehat{Y}^L(0;\alpha) - \widehat{Y}^L(1,\alpha)$ is a consistent estimator, and can be written as $g(\widehat{\boldsymbol{\mu}})$ for $g((x_1,x_2)^T) = x_1 - x_2$. From the Delta method, we know that

$$\sqrt{n}\left(\widehat{\overline{DE}}^L(\alpha) - \overline{DE}^L(\alpha)\right) \to N(0,\sigma^2)$$

for $\sigma^2 = \nabla g(\boldsymbol{\mu}_0)^T V(\boldsymbol{\mu}_0)\nabla g(\boldsymbol{\mu}_0)$, where $\nabla g((x_1,x_2)^T) = (\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2})^T = (1,-1)^T$, and $V(\boldsymbol{\mu}_0)$ is as in Theorem 1.

Proofs of asymptotic results for correctly specified propensity score

Lemma 1. *If condition 3 of Theorem 2 holds, then* $E\left[\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i, \boldsymbol{A}_i;\boldsymbol{\gamma}_0)\right] < \infty$

Proof of Lemma 1. Denote $\boldsymbol{\psi}_\gamma = (\psi_\gamma^1, \psi_\gamma^2, \ldots, \psi_\gamma^p)^T$. Then,

$$E_{F_0}^2\left(\psi_\gamma^k\right) \le E\left[\left(\psi_\gamma^k\right)^2\right] \le \sum_{l=1}^{p} E\left[\left(\psi_\gamma^l\right)^2\right] = E_{F_0}^2\left\|\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i, \boldsymbol{A}_i;\boldsymbol{\gamma})\right\|^2 < \infty \Rightarrow E_{F_0}\left(\psi_\gamma^k\right) < \infty$$

where the first inequality uses Jensen's inequality for $g(x) = x^2$. From this, we see that the score functions are integrable with finite expectation. □

Lemma 2. *Assuming that the conditions of Theorem 2 hold, the estimator* $\widehat{Y}_N^L(\alpha)$ *using the estimates of the correctly specified propensity score model is consistent for* $\boldsymbol{\mu}_0$.

Proof of Lemma 2. Consider the augmented estimated equations defined as $\Psi_n(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{L}_i; \boldsymbol{\theta})$, where

$$\boldsymbol{\psi}(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{L}_i; \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\psi}_\gamma(\boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\gamma}) \\ \psi_{0,\alpha}(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{L}_i; \mu^0, \boldsymbol{\gamma}) \\ \psi_{1,\alpha}(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{L}_i; \mu^1, \boldsymbol{\gamma}) \end{pmatrix}_{(p+2)\times 1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \mu^0, \mu^1)^T$ and p is the number of parameters of the parametric propensity score model. Note that $\psi_{a,\alpha}$ is now a function of $\boldsymbol{\gamma}$ since it uses the estimated propensity score. Denote the vector that solves $\Psi_n(\boldsymbol{\theta}) = 0$ as $\widehat{\boldsymbol{\theta}}$. The first p elements of $\widehat{\boldsymbol{\theta}}$ correspond to estimators of the propensity score model parameters, which are consistent for $\boldsymbol{\gamma}_0$. Since $\widehat{Y}_N^L(\alpha)$ based on the true propensity score is consistent, $f(\boldsymbol{a}_i|\boldsymbol{l}_i; \boldsymbol{\gamma})$ is differentiable in $\boldsymbol{\gamma}$ and therefore continuous, and $\boldsymbol{\gamma} \xrightarrow{p} \boldsymbol{\gamma}_0$, the last two elements of $\widehat{\boldsymbol{\theta}}$ which correspond to $\widehat{Y}^L(0,\alpha), \widehat{Y}^L(1,\alpha)$ using the estimated propensity score are consistent estimators of $\overline{Y}^L(0,\alpha), \overline{Y}^L(0,\alpha)$. □

Proof of Theorem 2. We will again use Theorem 5.41 of van der Vaart (1998). Since consistency has been established in Lemma 2, showing the four conditions stated in the proof of Theorem 1 for the augmented $\boldsymbol{\psi}$ will establish asymptotic normality. Denote $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}_0^T, \boldsymbol{\mu}_0^T)^T$.

Proof of (i). By the conditions of the theorem, $\boldsymbol{\gamma} \to \boldsymbol{\psi}_\gamma(\boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\gamma})$ is twice continuously differentiable. This implies that $\psi_{a,\alpha}(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \mu_0^a, \boldsymbol{\gamma})$, $a = 0, 1$ are three times continuously differentiable with respect to $\boldsymbol{\gamma}$. Therefore, the second order partial derivatives with respect to $\boldsymbol{\gamma}$ exist and are continuous. Moreover, since $\boldsymbol{\psi}_\gamma(\boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\gamma})$ is not a function of μ^a, and using (3.13), the second partial derivatives with respect to elements of $\boldsymbol{\mu} = (\mu^0, \mu^1)$ exist and are continuous. Lastly, all second order derivatives with respect to an element of $\boldsymbol{\mu}$ and an element of $\boldsymbol{\gamma}$ exist and are 0, and therefore continuous. This shows that $\boldsymbol{\theta} \to \boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i; \boldsymbol{\theta})$ is twice continuously differentiable.

Proof of (ii). We want to show that $E_{\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i} \|\psi(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\theta}_0)\|_2^2 < \infty$. But

$$E_{\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i} \|\psi(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\theta}_0)\|_2^2 = E_{\boldsymbol{L}_i, \boldsymbol{A}_i} \|\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i, \boldsymbol{A}_i; \boldsymbol{\gamma}_0)\|_2^2 + \sum_{a \in \{0,1\}} E_{\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i} \|\psi_{a,\alpha}(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{L}_i; \mu_0^a)\|^2,$$

where the first term is finite from the assumptions on the propensity score model, and the terms in the summation are finite from (3.12).

Proof of (iii). We want to show that the matrix $E_{\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i}\left[\dot{\psi}(\boldsymbol{y}_i,\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\theta}_0)\right]$ exists and is non singular. We have

$$\dot{\psi}(\boldsymbol{y}_i,\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial\boldsymbol{\gamma}^T}\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma})_{p\times p} & 0_{p\times 1} & 0_{p\times 1} \\ \frac{\partial}{\partial\boldsymbol{\gamma}^T}\boldsymbol{\psi}_{0,\alpha}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\mu^0,\boldsymbol{\gamma})_{1\times p} & -1 & 0 \\ \frac{\partial}{\partial\boldsymbol{\gamma}^T}\boldsymbol{\psi}_{1,\alpha}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\mu^1,\boldsymbol{\gamma})_{1\times p} & 0 & -1 \end{pmatrix},$$

where the the 0's in the top row are because $\boldsymbol{\psi}_\gamma$ is not a function of μ^0,μ^1. We have assumed that $E\left[\frac{\partial}{\partial\boldsymbol{\gamma}^T}\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma}_0)\right]$ exists and we will show that $E\left[\frac{\partial}{\partial\gamma^T}\psi_{a,\alpha}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\mu_0^a,\boldsymbol{\gamma}_0)\right]$ exists for $a=0,1$.

Showing that $E\left[\frac{\partial}{\partial\gamma^T}\psi_{a,\alpha}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\mu_0^a)\right]<\infty$ *for* $a=0,1$.

Note that even if the estimates of $\boldsymbol{\gamma}$ were used to define the counterfactual treatment allocation $P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)$, it is considered fixed as a function of $\boldsymbol{\gamma}$, since it is used to represent a fixed realistic treatment allocation program.

$$\begin{aligned}\frac{\partial}{\partial\gamma_k}\psi_{a,\alpha}(\boldsymbol{O}_i;\mu^a,\boldsymbol{\gamma}) &= \left(\frac{1}{n_i}\sum_{j=1}^{n_i}P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)I(A_{ij}=a)Y_{ij}\right)\left(\frac{\partial}{\partial\gamma_k}\frac{1}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}\right)\\ &= -\left(\frac{1}{n_i}\sum_{j=1}^{n_i}P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)I(A_{ij}=a)Y_{ij}\right)\left(\frac{\frac{\partial}{\partial\gamma_k}\log f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}\right)\\ &= -\psi_\gamma^k\left(\tilde{\boldsymbol{O}}_i;\boldsymbol{\gamma}\right)\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right),\end{aligned} \tag{3.14}$$

where $\psi_\gamma^k\left(\tilde{\boldsymbol{O}}_i;\boldsymbol{\gamma}\right)$ is the k^{th} component of $\boldsymbol{\psi}_\gamma\left(\tilde{\boldsymbol{O}}_i;\boldsymbol{\gamma}\right)$ for which $E_{F_0}\left[\psi_\gamma^k\left(\tilde{\boldsymbol{O}}_i;\boldsymbol{\gamma}_0\right)\right]<\infty$ (Lemma 1). Also, $\left|\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right|<M/\delta_o$ using the conditions of Theorem 1. So, we have shown that $E\left[\frac{\partial}{\partial\gamma^T}\psi_{a,\alpha}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\mu_0^a)\right]<\infty$.

From this, we conclude that $E_{F_0}\left[\dot{\psi}(\boldsymbol{y}_i,\boldsymbol{l}_i,\boldsymbol{a}_i;\boldsymbol{\theta})\right]$ exists. Furthermore, from the theorem assumptions we have that $E\left[\frac{\partial}{\partial\boldsymbol{\gamma}^T}\boldsymbol{\psi}_\gamma(\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\gamma}_0)\right]$ is non-singular and the rows of $\partial\boldsymbol{\psi}_\gamma/\partial\gamma^T$ are linearly independent. The bottom two rows are linearly independent to the rest since they are the only ones to include non-zero elements in the last two columns. From this, we conclude that the rows of $E\left[\dot{\psi}(\boldsymbol{Y}_i,\boldsymbol{L}_i,\boldsymbol{A}_i;\boldsymbol{\theta}_0)\right]$ are linearly independent, and the matrix is full rank and non-singular.

Proof of (iv). We need to show that $\exists$ integrable function $\alpha(\boldsymbol{o}_i)$ fixed, such that $\alpha(\boldsymbol{o}_i)$ dominates all the second order partial derivatives of $\boldsymbol{\psi}(\boldsymbol{o}_i;\boldsymbol{\theta})$. Therefore, we need to show that for $k,l \in \{1,2,\ldots,p\}$, $a \in \{0,1\}$:

1. $\left|\frac{\partial^2 \boldsymbol{\psi}_\gamma(\tilde{\boldsymbol{o}}_i;\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l}\right| \leq \alpha_{kl}(\boldsymbol{o}_i)$,
2. $\left|\frac{\partial^2 \boldsymbol{\psi}_\gamma(\tilde{\boldsymbol{o}}_i;\boldsymbol{\gamma})}{\partial\gamma_k\partial\mu^a}\right| \leq \alpha_k^a(\boldsymbol{o}_i)$,
3. $\left|\frac{\partial^2 \boldsymbol{\psi}_\gamma(\tilde{\boldsymbol{o}}_i;\boldsymbol{\gamma})}{\partial\mu^{a_1}\partial\mu^{a_2}}\right| \leq \alpha^{a_1a_2}(\boldsymbol{o}_i)$,
4. $\left|\frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{o}_i;\mu^a,\boldsymbol{\gamma})}{\partial\mu^{a_1}\partial\mu^{a_2}}\right| \leq \xi^{a_1a_2}(\boldsymbol{o}_i)$,
5. $\left|\frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{o}_i;\mu^a,\boldsymbol{\gamma})}{\partial\mu^{a_1}\partial\gamma_k}\right| \leq \xi_k^{a_1}(\boldsymbol{o}_i)$,
6. $\left|\frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{o}_i;\mu^a,\boldsymbol{\gamma})}{\partial\gamma_k\partial\gamma_l}\right| \leq \xi_{kl}(\boldsymbol{o}_i)$,

for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, where $\alpha_{kl}(\boldsymbol{o}_i), \alpha_k^a(\boldsymbol{o}_i), \alpha^{a_1a_2}(\boldsymbol{o}_i), \xi^{a_1a_2}(\boldsymbol{o}_i), \xi_k^a(\boldsymbol{o}_i), \xi_{kl}(\boldsymbol{o}_i)$ are F_0-integrable. If we show the above, by setting $\alpha(\boldsymbol{o}_i) = \max_{k,l,a}\{\alpha_{kl}(\boldsymbol{o}_i), \alpha_k^a(\boldsymbol{o}_i), \alpha^{a_1a_2}(\boldsymbol{o}_i), \xi^{a_1a_2}(\boldsymbol{o}_i), \xi_k^a(\boldsymbol{o}_i), \xi_{kl}(\boldsymbol{o}_i)\}$ we have that all second order partial derivatives are dominated by the F_0 integrable $\alpha(\boldsymbol{o}_i)$.

Since $\boldsymbol{\psi}_\gamma(\tilde{\boldsymbol{o}}_i;\boldsymbol{\gamma})$ is not a function of μ^a, conditions 2, 3 are easy to satisfy by setting $\alpha_k^a(\boldsymbol{o}_i) = \alpha^{a_1a_2}(\boldsymbol{o}_i) = 0$. The same is true for conditions 4, 5, since $\partial\psi_{a,\alpha}(\boldsymbol{o}_i;\mu_a,\boldsymbol{\gamma})/\partial\mu^{a_1} = -I(a=a_1)$ and therefore all second order derivatives that include at least one derivative with respect to μ^{a_1} will be equal to 0. So we can set $\xi^{a_1a_2}(\boldsymbol{o}_i) = \xi_k^a(\boldsymbol{o}_i) = 0$.

From the assumptions of the theorem, we know that $\exists\ \ddot{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i)$ integrable such that $\left|\frac{\partial^2 \boldsymbol{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i;\gamma)}{\partial\gamma_k\partial\gamma_l}\right| \leq \ddot{\psi}_\gamma(\boldsymbol{l}_i,\boldsymbol{a}_i)$, for all $\boldsymbol{\gamma}$ in a neighborhood of $\boldsymbol{\gamma}_0$. Then, $\alpha_{kl}(\boldsymbol{o}_i) = \ddot{\psi}_\gamma(\boldsymbol{o}_i)$ satisfy condition 1. Since γ_0 is in an open subset of the Euclidean space, there exists $\epsilon > 0$ such that the second partial derivatives of $\boldsymbol{\psi}_\gamma$ are dominated by $\ddot{\psi}$ for all $\boldsymbol{\gamma} \in \mathcal{N}^\epsilon(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| < \epsilon\}$, subset of the parameter space. Let $\overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \leq \epsilon/2\} \subset \mathcal{N}^\epsilon(\boldsymbol{\gamma}_0)$. Then, $\overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0)$ is a compact subset of the Euclidean space.

We will show that for $\boldsymbol{\gamma} \in \overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0)$ the second order partial derivatives in 6 are bounded by an integrable function. First, let's acquire their form:

$$\frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{O}_i; \mu^a, \boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} = \frac{\partial}{\partial \gamma_k} \left[-\psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij} \right) \right]$$

$$= - \frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij} \right)$$

$$- \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \frac{\partial}{\partial \gamma_k} \psi_{a,\alpha}(\boldsymbol{O}_i; \mu^a, \boldsymbol{\gamma})$$

$$= - \frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij} \right)$$

$$+ \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \psi_\gamma^k \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij} \right)$$

$$= \left[\psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \psi_\gamma^k \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) - \frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \right] \left[\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j} | A_{ij} = a, \boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i | \boldsymbol{L}_i)} I(A_{ij} = a) Y_{ij} \right]$$

where the first and third equation use (3.14), and the second equation is an application of the chain rule. Then

$$\left| \frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{O}_i; \mu^a, \boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} \right| < \frac{M}{\delta_o} \left| \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \psi_\gamma^k \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) - \frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \right| \tag{3.15}$$

For all $k, l \in \{1, 2, \ldots, p\}$, $\psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right)$, $\frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right)$ are differentiable and therefore continuous in $\boldsymbol{\gamma}$, implying that the function on the right-hand side of (3.15) is continuous in $\boldsymbol{\gamma}$.

Define $g(\boldsymbol{\gamma}) = E_{F_0} \left| \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \psi_\gamma^k \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) - \frac{\partial}{\partial \gamma_k} \psi_\gamma^l \left(\tilde{\boldsymbol{O}}_i; \boldsymbol{\gamma} \right) \right|$. Then $g(\boldsymbol{\gamma})$ is continuous in $\boldsymbol{\gamma}$. But since $\overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0)$ is a compact set, $g(\boldsymbol{\gamma})$ is bounded in $\overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0)$, and in fact achieves a maximum. Let

$$\xi_{kl}(\boldsymbol{o}_i) = \xi_{kl} = \frac{M}{\delta_o} \max \left\{ g(\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \overline{\mathcal{N}}^{\epsilon/2}(\boldsymbol{\gamma}_0) \right\}.$$

Then $\left| \frac{\partial^2 \psi_{a,\alpha}(\boldsymbol{O}_i; \mu^a, \boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} \right| < \xi_{kl}(\boldsymbol{o}_i), \ \forall \boldsymbol{\gamma} \in \overline{\mathcal{L}}^{\epsilon/2}(\boldsymbol{\gamma}_0)$, and ξ_{kl} is integrable since it is a constant function.

Then, set $\alpha(\boldsymbol{o}_i) = \max\{\alpha_{kl}(\boldsymbol{o}_i), \alpha_k^a(\boldsymbol{o}_i), \alpha^{a_1 a_2}(\boldsymbol{o}_i), \xi^{a_1 a_2}(\boldsymbol{o}_i), \xi_k^a(\boldsymbol{o}_i), \xi_{kl}(\boldsymbol{o}_i)\}$, and all second order partial derivatives are dominated by the F_0-integrable $\alpha(\boldsymbol{o}_i)$, for all $\boldsymbol{\theta} \in \mathcal{N}^{\epsilon/2}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\theta - \theta_0\| < \epsilon/2\}$.

From Theorem 5.41 of van der Vaart (1998), we have that

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \overset{N\to\infty}{\to} N(0, Q(\theta_0)),$$

where

$$Q(\theta_0) = A(\theta_0)^{-1}B(\theta)[A(\theta_0)^{-1}]^T$$

for $A(\theta_0) = E\left[\dot{\psi}(\boldsymbol{O}_i;\boldsymbol{\theta}_0)\right]$, and $B(\theta_0) = E\left[\psi(\boldsymbol{O}_i;\boldsymbol{\theta}_0)\psi(\boldsymbol{O}_i;\boldsymbol{\theta}_0)^T\right]$.
However, we are only interested in the bottom-right 2×2 submatrix of $Q(\theta_0)$ which corresponds to the asymptotic variance of $(\widehat{\mu}_0, \widehat{\mu}_1)^T$ when the propensity score model is estimated. Note that $A(\theta)$, $B(\theta)$ can be rewritten as

$$A(\theta) = \begin{bmatrix} A_{11} & 0 \\ A_{21} & -I_2 \end{bmatrix} \text{ where } A_{11} = E\left[\frac{\partial\psi_\gamma}{\partial\gamma^T}\right]_{p\times p} A_{21} = E\left[\frac{\partial(\psi_0,\psi_1)^T}{\partial\gamma^T}\right]_{2\times p},$$
$$B(\theta) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \text{ for } B_{11} = E\left[\psi_\gamma\psi_\gamma^T\right]_{p\times p} B_{12} = E[\psi_\gamma\psi_0, \psi_\gamma\psi_1]_{p\times 2},\ B_{21} = B_{12}^T, \text{ and}$$
$$B_{22} = E\begin{bmatrix} \psi_0^2 & \psi_0\psi_1 \\ \psi_1\psi_0 & \psi_1^2 \end{bmatrix},$$

where the arguments $(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i)$ have been suppressed. Then,

$$A(\theta)^{-1}B(\theta)[A(\theta)^{-1}]^T = \begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ A_{21}A_{11}^{-1} & -I_2 \end{bmatrix}\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}\begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ A_{21}A_{11}^{-1} & -I_2 \end{bmatrix}^T$$
$$= \begin{bmatrix} A_{11}^{-1}B_{11} & A_{11}^{-1}B_{12} \\ A_{21}A_{11}^{-1}B_{11} - B_{21} & A_{21}A_{11}^{-1}B_{12} - B_{22} \end{bmatrix}\begin{bmatrix} (A_{11}^{-1})^T & (A_{21}A_{11}^{-1})^T \\ \mathbf{0} & -I_2 \end{bmatrix}$$
$$= \begin{bmatrix} -I_p & -B_{11}^{-1}B_{12} \\ -A_{21} - B_{21} & -A_{21}B_{11}^{-1}B_{12} - B_{22} \end{bmatrix}\begin{bmatrix} -(B_{11}^{-1})^T & -(A_{21}B_{11}^{-1})^T \\ \mathbf{0} & -I_2 \end{bmatrix} \quad \text{(Since } A_{11} = -B_{11}\text{.)}$$
$$= \begin{bmatrix} \dots & \dots \\ \dots & (A_{21}+B_{21})B_{11}^{-1}A_{21}^T + A_{21}B_{11}^{-1}B_{12} + B_{22} \end{bmatrix} \quad (B_{11} \text{ symmetric} \Rightarrow B_{11}^{-1} \text{ symmetric})$$
$$= \begin{bmatrix} \dots & \dots \\ \dots & A_{21}B_{11}^{-1}A_{21}^T + A_{21}B_{11}^{-1}B_{12} + \left(A_{21}B_{11}^{-1}B_{12}\right)^T + B_{22} \end{bmatrix} \quad (B_{21} = B_{12}^T)$$

So the asymptotic covariance matrix of $(\widehat{\mu}_0, \widehat{\mu}_1)$ is equal to $A_{21}B_{11}^{-1}A_{21}^T + A_{21}B_{11}^{-1}B_{12} + \left(A_{21}B_{11}^{-1}B_{12}\right)^T + B_{22}$. □

3.8.4 Asymptotic variance of the population average potential outcome estimator

Denote $[V(\boldsymbol{\mu}_0)]_{ij}$ the ij element of the covariance matrix, and remember that $\mu_a = E_{F_0}[\overline{Y}_i(a,\alpha)]$. Then

$$
\begin{aligned}
[V(\boldsymbol{\mu}_0)]_{(a+1)(a+1)} =& E_{F_0}\left[\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}-\mu_a\right)^2\right] \\
=& E_{F_0}\left[\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right)^2\right]+\mu_a^2- \\
& -2\mu_a E_{F_0}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\right] \\
=& E_{F_0}\left[\overline{Y}_i^L(a,\alpha)^2\right]+\mu_a^2-2\mu_a E_{F_0}\left[\overline{Y}_i^L(a,\alpha)\right] \\
=& E_{F_0}\left[\overline{Y}_i^L(a,\alpha)^2\right]-E_{F_0}^2\left[\overline{Y}_i^L(a,\alpha)\right]=Var_{F_0}\left[\overline{Y}_i^L(a,\alpha)\right]
\end{aligned}
$$

$$
\begin{aligned}
[V(\boldsymbol{\mu}_0)]_{12} =& E_{F_0}\left[\left(\widehat{Y}_i^L(0,\alpha)-\mu_0\right)\left(\widehat{Y}_i^L(1,\alpha)-\mu_1\right)\right] \\
=& E_{F_0}\left[\widehat{Y}_i^L(0,\alpha)\widehat{Y}_i^L(1,\alpha)\right]-\mu_0 E_{F_0}\left[\widehat{Y}_i^L(1,\alpha)\right]-\mu_1 E_{F_0}\left[\widehat{Y}_i^L(0,\alpha)\right]+\mu_0\mu_1 \\
=& E_{F_0}\left[\widehat{Y}_i^L(0,\alpha)\widehat{Y}_i^L(1,\alpha)\right]-E_{F_0}\left[\widehat{Y}_i^L(0,\alpha)\right]E_{F_0}\left[\widehat{Y}_i^L(1,\alpha)\right] \\
=& Cov_{F_0}\left(\overline{Y}_i^L(0,\alpha),\overline{Y}_i^L(1,\alpha)\right)
\end{aligned}
$$

3.8.5 Population average potential outcome definitions in the literature

Assuming partial interference, Hudgens and Halloran (2008), and Tchetgen Tchetgen and VanderWeele (2012) defined the population average potential outcome as an average of the group-level potential outcomes $\overline{Y}(a;\alpha) = \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_i(a;\alpha)$. On the other hand, Liu et al. (2016) define the population average potential outcomes without assuming partial interference (and therefore without assuming the existence of interference clusters) as the average of the individual average potential outcomes. However, their asymptotic results are based on the assumption of partial interference, under which the population average potential outcome can be written as

$$
\overline{Y}^{Liu}(a;\alpha)=\frac{1}{\sum_{i=1}^{N}n_i}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\overline{Y}_{ij}(a;\alpha)=\sum_{i=1}^{N}\frac{n_i}{\sum_{i=1}^{N}n_i}\overline{Y}_i(a;\alpha).
$$

Therefore, the estimand of Liu et al. (2016), if partial interference is assumed, is equal to a weighted average of the group average potential outcomes with weights proportional to the number of individuals in the cluster.

Estimators for both quantities can be written in the form

$$\widehat{Y}(a;\alpha) = \sum_{i=1}^{N} \frac{d_i}{\sum_{i=1}^{N} d_i} \widehat{Y}_i(a;\alpha),\ d_i > 0, \tag{3.16}$$

where $\widehat{Y}_i(a;\alpha)$ is an unbiased estimator of the group average potential outcome for cluster i. The difference of the population average estimators lies in the specification of d_i, where $d_i = 1$ and $d_i = n_i$ accordingly, for the two definitions of population average potential outcome.

Proposition 1. *Under the assumption of partial interference (which is also assumed by Liu et al. (2016) in their asymptotic results), all population average potential outcome estimators of the form (3.16) for which $d_i > 0$ does not depend on N, $E_{F_0}[d_i] < \infty$, and $d_i \amalg \overline{Y}_i(a;\alpha)$ are consistent for $E_{F_0}\left[\overline{Y}_i(a;\alpha)\right]$.*

Proof of Proposition 1. This can be shown by considering the estimating equation $\sum_{i=1}^{N} G_i(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i, d_i; \mu) = 0$, where

$$G_i(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i, d_i; \mu) = d_i \left[\widehat{Y}_i(a;\alpha) - \mu\right],$$

The solution to this equation is

$$\widehat{\mu} = \sum_{i=1}^{N} \frac{d_i}{\sum_{i=1}^{N} d_i} \widehat{Y}_i(a,\alpha),$$

and the solution to $\int G_i(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i, d_i; \mu) \mathrm{d}F_0(\boldsymbol{y}_i, \boldsymbol{l}_i, \boldsymbol{a}_i) = 0$ is

$$\begin{aligned}\mu_0 =& \frac{E[d_i \widehat{Y}_i(a;\alpha)]}{E[d_i]} = \frac{E\left\{E\left[d_i \widehat{Y}_i(a;\alpha) | d_i\right]\right\}}{E[d_i]} = \frac{E\left\{d_i E\left[\widehat{Y}_i(a;\alpha)\right]\right\}}{E[d_i]} \\ =& \frac{E\left[d_i \overline{Y}_i(a;\alpha)\right]}{E[d_i]} = E_{F_0}\left[\overline{Y}_i(a;\alpha)\right],\end{aligned}$$

since $d_i \amalg \overline{Y}_i(a;\alpha)$.

Since G_i is monotone in μ, both $\sum_{i=1}^{N} G_i$ and $\int G_i$ are monotone in μ which implies uniqueness of the roots and establishes $\widehat{\mu} \xrightarrow{p} \mu_0$. □

Based on this, assuming $n_i \amalg \overline{Y}_i(a;\alpha)$ both estimators are consistent for the same quantity. However, when the propensity score is known, the weighting scheme $d_i = c$, constant, leads to the asymptotically most efficient estimator among all of the estimators of the form (3.16), based on the following proposition. Since two estimators using d_i and $d_i' = cd_i$ are exactly the same, the estimator (3.16) for $d_i = 1$ is the asymptotically efficient estimator.

Proposition 2. *Assuming that the conditions of Theorem 1 and Proposition 1 hold, and $\exists M_d$ such that $d_i < M_d, \forall i$, then $\widehat{Y}(a;\alpha) = \frac{1}{N}\sum_{i=1}^{N}\widehat{Y}_i(a;\alpha)$ is the asymptotically most efficient estimator of $E_{F_0}\left[\overline{Y}_i(a;\alpha)\right]$ among all estimators of the class (3.16).*

Proof of Proposition 2. Based on Proposition 1, $\widehat{\mu}_{\boldsymbol{d}}(a;\alpha) = \sum_{i=1}^{N}\frac{d_i}{\sum d_i}\widehat{Y}_i(a;\alpha)$ are consistent for $\mu_0 = E_{F_0}\left[\overline{Y}_i(a;\alpha)\right]$. Since $G_i(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i, d_i;\mu) = d_i\left[\widehat{Y}_i(a;\alpha) - \mu\right]$ is monotone decreasing in μ with $\frac{\partial}{\partial\mu}G_i(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i, d_i;\mu) = -d_i < 0$, we have that $\mu_{\boldsymbol{d}}$ and μ_0 are isolated roots of $\sum_{i=1}^{N} G_i = 0$ and $\int G_i = 0$. Also, $E[\frac{\partial}{\partial\mu}G_i(\boldsymbol{Y}_i, \boldsymbol{L}_i, \boldsymbol{A}_i, d_i;\mu)] = -E[d_i] \neq 0$. Lastly, from (3.12) and $d_i < M_d$ we have that $\int G_i^2$ is bounded by $M_d^2c^2$. We can straightforwardly use M-estimation theory to acquire the asymptotic variance. Lemma A in section 7.2.1 of Serfling (1980), $\sqrt{n}(\widehat{\mu}_{\boldsymbol{d}} - \mu_0) \xrightarrow{d} N(0, \sigma^2(\boldsymbol{d}))$, where $\sigma^2(\boldsymbol{d}) = E_{F_0}[G_i^2(\cdot, d_i;\mu_0)]/E_{F_0}^2[d_i]$.
We will show that $\sigma^2(\boldsymbol{d})$ is minimized when $d_i = 1, \forall i$. Since $d_i \amalg \overline{Y}_i(a;\alpha)$, we have that $d_i^2 \amalg (\overline{Y}_i(a;\alpha) - \mu_0)^2$. Then $\sigma_{\boldsymbol{d}}^2 = \dfrac{E_{F_0}\left[d_i^2(\widehat{Y}(a;\alpha) - \mu_0)^2\right]}{[E_{F_0}d_i]^2} = \dfrac{E_{F_0}[d_i^2]}{E_{F_0}^2[d_i]}E_{F_0}\left[(\widehat{Y}_i(a;\alpha) - \mu_0)^2\right] = \dfrac{E_{F_0}[d_i^2]}{E_{F_0}^2[d_i]}\sigma_{\boldsymbol{1}}^2$, where $\sigma_{\boldsymbol{1}}^2$ is the asymptotic variance of the estimator for $d_i = 1$. From Jensen's inequality, and since $\phi(x) = x^2$ is a convex function, we have that $E_{F_0}^2[d_i] \leq E_{F_0}[d_i^2]$, which establishes $\sigma_{\boldsymbol{d}}^2 \geq \sigma_{\boldsymbol{1}}^2$. Equality holds if and only if all values d_i are equal. □

3.8.6 Calculating cluster-intercept for a specific cluster average propensity of treatment

As described in section 3.5.3, ξ_i^{α} is chosen such that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} P_{\alpha,L}(A_{ij} = 1|\boldsymbol{L}_i) = \alpha, \tag{3.17}$$

where $\text{logit}P_{\alpha,L}(A_{ij}=1|\boldsymbol{L}_i)=\xi_i^\alpha+L_{ij}\boldsymbol{\delta}$, and $L_{ij}=(L_{1ij},L_{2ij},\ldots,L_{pij})^T$ the value of the p predictors of the propensity score model. Then, (3.17) can be rewritten as

$$\frac{1}{n_i}\sum_{j=1}^{n_i}P_{\alpha,L}(A_{ij}=1|\boldsymbol{L}_i)=\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\exp\{\xi_i^\alpha+L_{ij}\boldsymbol{\delta}\}}{1+\exp\{\xi_i^\alpha+L_{ij}\boldsymbol{\delta}\}}=\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\exp\{L_{ij}\boldsymbol{\delta}\}}{\exp\{\xi_i^\alpha\}+\exp\{L_{ij}\boldsymbol{\delta}\}}=\alpha$$

$$\iff\left|\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\exp\{L_{ij}\boldsymbol{\delta}\}}{\exp\{-\xi_i^\alpha\}+\exp\{L_{ij}\boldsymbol{\delta}\}}-\alpha\right|=0$$

Since the only unknown is ξ_i^α, we use optimization techniques and set ξ_i^α to be the value ξ at which the function

$$g(\xi)=\left|\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\exp\{L_{ij}\boldsymbol{\delta}\}}{\exp\{-\xi\}+\exp\{L_{ij}\boldsymbol{\delta}\}}-\alpha\right|$$

is minimized.

Product of small numbers

Calculating the estimator (3.4) above can lead to problems since it includes (both in the numerator and the denominator) products of probabilities. This can be simplified by rewriting the estimator as:

$$\begin{aligned}\widehat{Y}_i(a;\alpha)=&\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}I(A_{ij}=a)Y_{ij}\\=&\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{\pi(\boldsymbol{A}_{i,-j};\alpha)}\frac{\pi(\boldsymbol{A}_i;\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}\frac{I(A_{ij}=a)Y_{ij}}{\alpha^a(1-\alpha)^{1-a}},\end{aligned}$$

where $\pi(\boldsymbol{a}_i;\alpha)=\prod_{j=1}^{n_i}\alpha^{a_{ij}}(1-\alpha)^{1-a_{ij}}$, and

$$f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)=\int\prod_{j=1}^{n_i}f_{\boldsymbol{A}|\boldsymbol{L},i}(A_{ij}|L_{ij},\tilde{\boldsymbol{\delta}},\sigma_b^2,b_i)f(b_i;\sigma_b^2)\mathrm{d}b_i.$$

For the first ratio, we write

$$\frac{P_{\alpha,L}(\boldsymbol{A}_{i,-j}|A_{ij}=a,\boldsymbol{L}_i)}{\pi(\boldsymbol{A}_{i,-j};\alpha)}=\prod_{j'\neq j}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i,\alpha}(A_{ij}|L_{ij},\alpha)}{\alpha^{A_{ij'}}(1-\alpha)^{1-A_{ij'}}},$$

and for the second ratio

$$\begin{aligned}\frac{\pi(\boldsymbol{A}_i;\alpha)}{f_{\boldsymbol{A}|\boldsymbol{L},i}(\boldsymbol{A}_i|\boldsymbol{L}_i)}&=\left[\frac{\int\prod_{j=1}^{n_i}f_{\boldsymbol{A}|\boldsymbol{L},i}(A_{ij}|L_{ij},\tilde{\boldsymbol{\delta}},\sigma_b^2,b_i)f(b_i;\sigma_b^2)\mathrm{d}b_i}{\prod_{j=1}^{n_i}\alpha^{A_{ij}}(1-\alpha)^{1-A_{ij}}}\right]^{-1}\\&=\left[\int\prod_{j=1}^{n_i}\frac{f_{\boldsymbol{A}|\boldsymbol{L},i}(A_{ij}|L_{ij},\tilde{\boldsymbol{\delta}},\sigma_b^2,b_i)}{\alpha^{A_{ij}}(1-\alpha)^{1-A_{ij}}}f(b_i;\sigma_b^2)\mathrm{d}b_i\right]^{-1},\end{aligned}$$

which is much more stable to calculate.

References

ALI, M., EMCH, M., VON SEIDLEIN, L., YUNUS, M., SACK, D. A., RAO, M., HOLMGREN, J. and CLEMENS, J. D. (2005). Herd immunity conferred by killed oral cholera vaccines in Bangladesh: A reanalysis. *Lancet* **366** 44–49.

ALLEN, J. (2002). Chemistry in the Sunlight. *Earth Observatory NASA* .

ANTONELLI, J., MAZUMDAR, M., BELLINGER, D., CHRISTIANI, D. C., WRIGHT, R. and COULL, B. A. (2017a). Bayesian variable selection for multi-dimensional semiparametric regression models .
URL `https://arxiv.org/pdf/1711.11239.pdf`

ANTONELLI, J., ZIGLER, C. and DOMINICI, F. (2017b). Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics* **18** 553–568.

BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* **17** 1103–20.

BARKLEY, B. G., HUDGENS, M. G., CLEMENS, J. D., ALI, M. and EMCH, M. E. (2017). Causal Inference from Observational Studies with Clustered Interference .
URL `https://arxiv.org/pdf/1711.04834.pdf`

BELL, M. L., MCDERMOTT, A., ZEGER, S. L., SAMET, J. M. and DOMINICI, F. (2004). Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000. *JAMA* **292** 2372–2378.

BELL, M. L., PENG, R. D. and DOMINICI, F. (2006). The exposure-response curve for

ozone and risk of mortality and the adequacy of current ozone regulations. *Environmental health perspectives* **114** 532–6.

BIA, M., FLORES, C. A., FLORES-LAGUNES, A. and MATTEI, A. (2014). A Stata package for the application of semiparametric estimators of doseâĂŞresponse functions. *The Stata Journal* **14** 580–604.

CHANG, H. H., REICH, B. J. and MIRANDA, M. L. (2013). A spatial time-to-event approach for estimating associations between air pollution and preterm birth. *Journal of the Royal Statistical Society. Series C, Applied statistics* **62** 167–179.

CONGDON, P. (2013). Assessing the impact of socioeconomic variables on small area variations in suicide outcomes in england. *International Journal of Environmental Research and Public Health* **10** 158–177.

DANIELS, M. J., DOMINICI, F., SAMET, J. M. and ZEGER, S. L. (2000). Estimating Particulate Matter-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest US Cities. *American Journal of Epidemiology* **152** 397–406.

DANIELS, M. J., DOMINICI, F., ZEGER, S. L. and SAMET, J. M. (2004). The National Morbidity, Mortality, and Air Pollution Study. Part III: PM10 concentration-response curves and thresholds for the 20 largest US cities. *Research report (Health Effects Institute)* 1–21; discussion 23–30.

DOMINICI, F., DANIELS, M., ZEGER, S. L. and SAMET, J. M. (2002). Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships. *Journal of the American Statistical Association* **97** 100–111.

EFTIM, S. E., SAMET, J. M., JANES, H., MCDERMOTT, A. and DOMINICI, F. (2008). Fine Particulate Matter and Mortality: A Comparison of the Six Cities and American Cancer Society Cohorts With a Medicare Cohort. *Epidemiology* **19** 209–216.

FERRACCI, M., JOLIVET, G. and VAN DEN BERG, G. J. (2014). Evidence of Treatment Spillovers Within Markets. *Review of Economics and Statistics* **96** 812–823.

FINLEY, A. O., BANERJEE, S. and CARLIN, B. P. (2007). spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models. *Journal of statistical software* **19** 1–24.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29** 1189–1232.

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016.

GELMAN, A. and RUBIN, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7** 457–511.

GOVINDARAJULU, U. S., MALLOY, E. J., GANGULI, B., SPIEGELMAN, D. and EISEN, E. A. (2009). The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *The international journal of biostatistics* **5** Article 2.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrica* **82** 711–732.

GU, X. S. and ROSENBAUM, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Source Journal of Computational and Graphical Statistics* **2** 405–420.

HASTIE, T. (2017). gam: Generalized Additive Models.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized Additive Models. *Statistical Science* **1** 297–318.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B* **55** 757–796.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**.

HIRANO, K. and IMBENS, G. W. (2004). The Propensity Score with Continuous Treatments * .

HO, C., DANIEL, E., IMAI, G., KING, E. and STUART (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* **15** 199–236.

HODGES, J. S. and REICH, B. J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician* **64** 325–334.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward Causal Inference With Interference. *Journal of the American Statistical Association* **103** 832–842.

IVERSON, H. K. and RANDLES, R. H. (1989). The Effects on Convergence of Substituting Parameter Estimates into U-Statistics and Other Families of Statistics. *Probability Theory and Related Fields* **81** 453–471.

JERRETT, M., BURNETT, R. T., POPE III, C. A., ITO, K., THURSTON, G., KREWSKI, D., SHI, Y., CALLE, E. and THUN, M. (2009). Long-Term Ozone Exposure and Mortality. *N Engl J Med* **360** 1085–95.

KEELE, L., TITIUNIK, R. and ZUBIZARRETA, J. (2015). Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout. *Journal of Royal Statistical Society A* **178** 223–239.

KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1229–1245.

KING, G. and NIELSEN, R. (2016). Why Propensity Scores Should Not Be Used for Matching .
URL `https://gking.harvard.edu/files/gking/files/psnot.pdf`

LEE, D. and NEOCLEOUS, T. (2010). Bayesian quantile regression for count data with

application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59** 905–920.

LEE, D. and SARRAN, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics* **26** 477–487.

LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association* **109** 288–301.

LIU, L., HUDGENS, M. G. and BECKER-DREPS, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika* **103** 829–842.

LUNA, X. D., WAERNBAUM, I. and RICHARDSON, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98** 861–875.

MADIGAN, D., YORK, J. and ALLARD, D. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review* **63** 215–232.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21** 1087–1092.

MINASNY, B. and MCBRATNEY, A. B. (2005). The Matérn function as a general model for soil variograms. *Geoderma* **128** 192–207.

NEYMAN, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* **5** 465–480.

PACIOREK, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science* **25** 107–125.

PAPADOGEORGOU, G., CHOIRAT, C. and ZIGLER, C. M. (2018). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics* **00** 1–17.

PAPADOGEORGOU, G., MEALLI, F. and ZIGLER, C. (2017). Causal inference for interfering units with cluster and population level treatment allocation programs .
URL https://arxiv.org/pdf/1711.01280.pdf

PEREZ-HEYDRICH, C., HUDGENS, M. G., HALLORAN, M. E., CLEMENS, J. D., ALI, M. and EMCH, M. E. (2015). Assessing Effects of Cholera Vaccination in the Presence of Interference. *Biometrics* **33** 395–401.

RAFTERY, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology* **25** 111–163.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92** 179–191.

RIDGEWAY, G. (2007). Generalized Boosted Models: A guide to the gbm package .

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70** 41–55.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

RUBIN, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Source Journal of the American Statistical Association* **75** 591–593.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2** 808–840.

SCHAFER, J. (2015). causaldrf: Tools for Estimating Causal Dose Response Functions.

SCHOLZE, M., BOEDEKER, W., FAUST, M., BACKHAUS, T., ALTENBURGER, R. and HORST, L. (2001). A general best-ft method for concentration-response curves and the estimation of low-effect concentrations. *Environmental Toxicology and Chemistry* **20** 448–457.

SCHWARTZ, J., LADEN, F. and ZANOBETTI, A. (2002). The Concentration-Response Relation between PM2.5 and Daily Deaths. *Environmental Health Perspectives* **110** 1025–1029.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

SHADDICK, G., LEE, D., ZIDEK, J. V. and SALWAY, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Annals of Applied Statistics* **2** 1249–1270.

SHI, L., ZANOBETTI, A., KLOOG, I., COULL, B. A., KOUTRAKIS, P., MELLY, S. J. and SCHWARTZ, J. D. (2016). Low-Concentration PM2.5 and Mortality: Estimating Acute and Chronic Effects in a Population-Based Study. *Environmental health perspectives* **124** 46–52.

SOBEL, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate? *Journal of the American Statistical Association* **101** 1398–1407.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25** 1–21.

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research* **21** 55–75.

VAN DER VAART, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

VANSTEELANDT, S., BEKAERT, M. and CLAESKENS, G. (2012). On model selection and model misspecification in causal inference. *Statistical methods in medical research* **21** 7–30.

VERBITSKY-SAVITZ, N. and RAUDENBUSH, S. W. (2012). Causal Inference Under Interference in Spatial Settings : A Case Study Evaluating Community Policing Program in Chicago. *Epidemiologic Methods* **1** 105–130.

WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* **71** 654–665.

WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian Effect Estimation Accounting for Adjustment Uncertainty. *Biometrics* **68** 661–671.

WARD, J. H. J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58** 236–244.

WATANABE, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research* **11** 3571–3594.

WILSON, A. and REICH, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics* **70** 852–861.

ZANOBETTI, A. and SCHWARTZ, J. (2007). Particulate air pollution, progression, and survival after myocardial infarction. *Environmental health perspectives* **115** 769–75.

ZEGER, S. L., DOMINICI, F., MCDERMOTT, A. and SAMET, J. M. (2008). Mortality in the Medicare population and chronic exposure to fine particulate air pollution in urban centers (2000-2005). *Environmental health perspectives* **116** 1614–9.

ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289–302.

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Casual Inference Methods in Air Pollution Research"

presented by Georgia Papadogeorgou

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature .....

Typed name: Prof. Francesca Dominici

Signature .....

Typed name: Prof. Corwin Zigler

Signature .....

Typed name: Prof. Giovanni Parmigiani

Signature .....

Typed name:

Date: April 26, 2018